An improved probabilistic account of counterfactual reasoning

An improved probabilistic account of counterfactual reasoning

Christopher G. Lucas
University of Edinburgh

Charles Kemp
Carnegie Mellon University

When people want to identify the causes of an event, assign credit or blame, or learn from their mistakes, they often reflect on how things could have gone differently. In this kind of reasoning, one considers a counterfactual world in which some events are different from their real-world counterparts and considers what else would have changed. Researchers have recently proposed several probabilistic models that aim to capture how people do (or should) reason about counterfactuals. We present a new model and show that it accounts better for human inferences than several alternative models. Our model builds on the work of Pearl (2000), and extends his approach in a way that accommodates backtracking inferences and that acknowledges the difference between counterfactual interventions and counterfactual observations. We present six new experiments and analyze data from four experiments carried out by Rips (2010), and the results suggest that the new model provides an accurate account of both mean human judgments and the judgments of individuals.

In addition to reasoning about actual states of affairs, humans often reason about what might have been. A doctor might ask "if Alice had not been treated with the experimental drug, would she have survived?" and a parent might tell a child that "if you had been paying attention, you wouldn't have gotten hurt." As these examples suggest, counterfactual reasoning helps us to evaluate our past decisions, and to learn what mistakes to avoid in the future (Mandel & Lehman, 1996; Zeelenberg, van Dijk, Manstead, & van der Pligt, 2000; Spellman, Kincannon, & Stose, 2005; Epstude & Roese, 2008). The answers to counterfactual questions also allow us to assign blame for harmful outcomes, and to give credit for beneficial outcomes (Spellman & Kincannon, 2001).

There has been a great deal of research into the factors that influence counterfactual judgments, and several formal models of counterfactual reasoning have been proposed. Of these, Pearl's (2000, 2013) account has been especially influential. Pearl's model provides a clear account of how information about causal structure and probability should influence counterfactual judgments, but falls short as a descriptive account of human behavior in certain contexts. This paper describes a new model that improves on Pearl's as an account of human counterfactual reasoning.

The issues that we consider can be introduced using a simple running example. Imagine that cooking some bacon activates a smoke alarm, which in turn disturbs the neighbors. Let B denote cooking bacon, S denote the smoke alarm activating, and N denote the neighbors being disturbed, as depicted in the graph in Figure 1a. All three variables are binary variables that can be true (t) or false (f). Following Pearl (2000) and others (Goldvarg & Johnson-Laird, 2001; Luhmann & Ahn, 2005; Schölkopf et al., 2012), our approach assumes that causal relationships are intrinsically deterministic and that apparently unreliable or stochastic relationships are influenced by unobserved factors. For example, unobserved factors such as the direction of air currents may help to determine whether cooking bacon triggers the smoke alarm on any given occasion. Under this view, causal systems can be viewed as collections of effects, where each effect is a deterministic function of its known causes and zero or more hidden variables. The hidden variables are exogenous: that is, they are not effects of any of the observed variables. A causal system expressed in this way is called a *functional causal model*. Figure 1b shows an example in which exogenous variables U_B, U_S and U_N have been added to the system in Figure 1a to capture background conditions that influence the causal links in our example. Figure 1b indicates that U_B determines whether or not bacon is cooked, that U_S and B jointly determine whether the smoke alarm activates, and that U_N and S jointly determine whether the neighbors are disturbed.

Suppose that B, S, and N are all known to be true: we cooked some bacon, the alarm activated, and the neighbors were disturbed. We will consider counterfactual questions such as "if the smoke alarm had not activated, would bacon still have been cooked?" (Figure 1). Our approach maintains a functional causal model for the actual world in which B, S

A preliminary version of this work was presented at the 34th Annual Meeting of the Cognitive Science Society. We thank David Danks and David Over for valuable comments on this research and on this manuscript. This work was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative and by NSF award CDI-0835797.

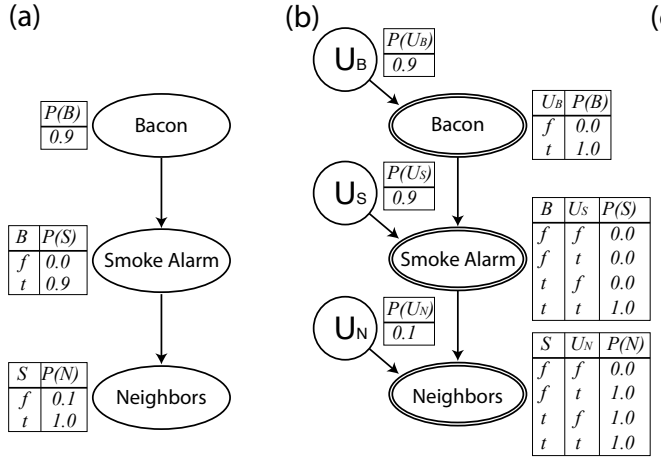


Figure 1. A causal graphical model and a corresponding functional causal model. The conditional probability tables at each node give the probability that a variable is present (t) or absent (f), conditional on its causes. (a) A causal graphical model in which each relationship is a noisy effect of its parents. (b) A functional causal model that introduces exogenous variables to the model in (a) and assumes that each variable is a deterministic function of its parents, denoted by double boundaries on the nodes. (c) A table illustrating a counterfactual query. Even though all three variables were present in the real-world, we might reason about a counterfactual world in which the smoke alarm had not sounded.

and N are all set to true, and a second model for the counterfactual world in which S is set to false. The graph structures and the probability distributions for the two models are identical, but the variables in the two models may take different values. To specify how the models are related, our approach makes use of the notion of stability. When stability is high, the exogenous variables in the two models are likely to take identical values. For example, when stability is high, the air currents (variable U_S) are expected to be identical across the real and counterfactual worlds. When stability is low, the exogenous variables are generated independently for the two models, and may therefore take different values.

The counterfactual model supports two different methods of reasoning about the counterfactual scenario in Figure 1. The first method treats the counterfactual premise as an intervention. For example, we might imagine a counterfactual world in which somebody directly intervened on the smoke alarm to disable it. The second method treats the counterfactual premise as an observation. For example, we might imagine that we had simply observed the alarm to be inactive. The two methods can lead to different conclusions. For example, if reasoning about the counterfactual intervention, we might conclude that bacon would still have been cooked on that day, and that the alarm failed to activate only because it had been disabled. If reasoning about the counterfactual observation, we might conclude that bacon was not cooked in the counterfactual world, because otherwise there would be no good explanation of the alarm's failure to activate. Sloman and Lagnado (2005) have previously shown that people distinguish between counterfactual interventions and counterfactual observations, but ours is the first psychological model that highlights this distinction.

Because our approach builds on the structural model of counterfactual reasoning developed by Pearl (2000, 2013), we refer to it as the *extended structural model* or ESM for short. Pearl's approach can be viewed as a special case of the ESM in which stability is maximal, and in which counterfactual premises are interpreted as interventions. We will argue, however, that this special case is not enough to account for human counterfactual reasoning. In particular, we will show how our approach goes beyond Pearl's by accounting for *backtracking* counterfactuals, such as the inference that bacon would not have been cooked if the alarm had failed to activate. Given the information in Figure 1c, the ESM allows for the possibility that variables upstream of S might explain the counterfactual premise that S is false, and can therefore infer that B is likely to be false. Pearl's approach, in contrast, predicts that upstream variables will be unchanged.

Although the ESM relies on functional causal models, there are alternative accounts of counterfactual reasoning that allow causal relationships to be intrinsically stochastic (Hiddleston, 2005; Rips, 2010; Dehghani, Iliev, & Kaufmann, 2012). Of these accounts, the framework that is best developed as a psychological model was proposed and evaluated by Rips (2010). One strength of Rips' account is that it is able to account for backtracking counterfactuals. We present several experiments, however, which suggest that the ESM performs better than Rips' model as an account of human reasoning.

The counterfactual inferences that we consider correspond

to inferences about conditionals of the form "if the alarm had not activated, then bacon would still have been cooked." Our work therefore contributes to an extensive literature that has explored how people reason about conditional statements (Stalnaker, 1981; Edgington, 1995; Evans & Over, 2004). The ESM is related most directly to previous accounts of conditional reasoning that rely on probabilistic inference (Oaksford & Chater, 2007; Evans & Over, 2004; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Singmann, Klauer, & Over, 2014). Among these accounts are some that focus specifically on probabilistic inference over causal models (Ali, Chater, & Oaksford, 2011; Fernbach & Erb, 2013). All of the models compared in this paper can therefore be viewed as attempts to extend the causal approach to conditional reasoning in a way that supports inferences about counterfactual conditionals.

In the next section, we summarize previous causal models of counterfactual reasoning and introduce our new approach in detail. We then present a series of experiments that test our model's core commitments and compare it against several alternatives.

Theories of counterfactual reasoning

Counterfactual reasoning has been extensively discussed by philosophers, linguists, and psychologists, and we will not be able to do justice to all of the work in this area. Instead, we begin by discussing some ideas that are widely shared by accounts of counterfactual reasoning, and then focus in detail on several recent theories that make use of causal Bayes nets.

Suppose that an individual is asked to evaluate an argument of the form "If P, then Q." Many theorists propose that the individual does so by (1) adding P hypothetically to her set of beliefs, (2) making some adjustments so that the resulting set is coherent, then (3) assessing Q in light of this adjusted set of beliefs (Oaksford & Chater, 2010a; Unterhuber, 2013). This three step procedure is often called the Ramsey test, and it lies at the heart of many theories of conditional reasoning in general and of counterfactual reasoning in particular. All of the theories that we will consider are broadly consistent with the Ramsey test.

Perhaps the most influential account of counterfactual reasoning makes use of a similarity measure over possible worlds (Stalnaker, 1981; Lewis, 1973b). Roughly speaking, this account proposes that $P \rightarrow Q$ is valid if Q is true in all of the worlds that make P true and are as as similar as possible to the actual world. The possible worlds account includes room for many different theories that define similarity in different ways. For example, Pearl (2000) showed that his account of counterfactual reasoning is a special case of the possible worlds approach if similarity is defined in a certain way.

As we describe later, our new theory can also be formulated in terms of possible worlds. Unlike most previous theories, however, we propose that counterfactual inferences are generated by reasoning about a diverse set of possible worlds, including worlds that are not as similar as possible to the actual world. Our approach may therefore seem incompatible with stage (2) of the Ramsey test, at least as this test is often described. For example, Harper (1981, p. 5) writes that the individual should make the "minimal revision" required to assume the antecedent P, and Bennett (2003, p. 29) suggests that the individual should adjust her belief system "in the most natural, conservative manner." For a possible worlds account, this minimality assumption corresponds to the idea that counterfactuals should be evaluated by considering only possible worlds that differ in minimal respects from the actual world. Although this minimality assumption is sometimes taken for granted, it does not appear in the footnote that is the original source of the Ramsey test (Ramsey, 1978), and therefore does not seem to be an essential component of this test.

Although the minimality assumption is extremely common, philosophers have offered some reasons to question it (Bennett, 2003; McDermott, 2007). Several authors point out that the assumption may not be appropriate when reasoning about ordinal variables (Jackson, 1977; Slote, 1978). For example, Bennett (2003, p. 180) gives the example of an individual with four siblings who is evaluating the argument "if I had more siblings than I do, I would have had exactly five." The minimality assumption suggests that the argument is valid, but it seems reasonable for the individual to think about possible worlds in which he had five, six, and perhaps even more siblings. Philosophers have also suggested that the assumption may not be appropriate when reasoning about stochastic events (Walters, 2009; Nozick, 1981). This prior philosophical work provides some justification for considering a theory that abandons the minimality assumption, but the ultimate test of our new theory will be whether or not it accounts for empirical data. In later sections we describe experiments designed to compare our theory with several prominent alternatives, and argue that the results raise serious challenges for any theory that relies on the minimality assumption.

Causal Bayes nets

The theories that we will compare most extensively are all formulated in terms of Causal Bayes nets, or CBNs (Pearl, 2000). This section reviews the properties of CBNs, and the following sections explain how CBNs can be used to develop computational accounts of counterfactual reasoning.

CBNs were originally proposed as a framework for understanding causal systems and have their origins in Bayes nets, a formalism that is useful for efficiently representing and reasoning about stochastic systems. A causal Bayes net, such as the one shown in Figure 1a, includes a set of nodes corresponding to variables and a set of directed edges — de-

picted as arrows between nodes — that capture the causal dependencies between variables. Every arrow in Figure 1a represents a direct causal relationship between one of the nodes, e.g., B is a direct cause of S. To complete a CBN one must specify precisely how each variable is influenced by its direct causes. For discrete variables, this information can be represented using conditional probability tables like those under each of the nodes in Figure 1a. These tables can capture the form of the relationship between variables, as in the tables under nodes S and N, or how common a particular state is, as in the table under B.

Causal Bayes nets support two basic operations, conditioning and intervention, which correspond to different ways in which a variable can take a particular value. If we observe that a variable has taken a particular value, e.g. $S = f$, we can predict the values of other variables by computing their probability distributions conditional on the observation. On the other hand, if we are reasoning about the effects of experiments or outside influences on a system, this sort of reasoning is not warranted: if we alter variable S, there will be consequences for its effect N, but not its cause B. In a causal Bayes net, an intervention detaches the target variable from its causes, removing the relevant edges, and only then do we compute the conditional distributions for the remaining variables.

Some of the models we will discuss use a formalism related to causal Bayes nets, called functional causal models (FCMs) by Pearl (2000). FCMs differ from causal Bayes nets in that they do not treat causal relationships as inherently stochastic, but instead represent noise and uncertainty using hidden, exogenous variables. For example, the CBN in Figure 1a shows stochastic relationships between B and S and between S and N. The FCM in Figure 1b introduces three exogenous variables U_B, U_S, and U_N, and indicates that the endogenous variables B, S, and N follow deterministically from their parents (denoted by double-edged boundaries). Like Figure 1a, each variable has an associated probability table, but the only probabilities that differ from 0 or 1 are the rates of the three exogenous variables.

The FCM in Figure 1b and the CBN in Figure 1a induce the same joint distribution on variables B, S, and N, as can be shown by marginalizing over the exogenous variables. As a result, the two models give the same answers to questions like "what is the probability of S if B is present?" Similarly, the models respond identically to causal interventions on endogenous variables, which can never influence the exogenous variables. Despite these similarities, Pearl has argued that FCMs should be preferred to CBNs, in part because FCMs provide a more complete account of counterfactual reasoning. We will assess this argument in later sections by comparing models that rely on FCMs with models that rely on CBNs. The difference between FCMs and CBNs is directly relevant to the long-running debate about whether causation is truly deterministic (Hume, 1748; Reichenbach, 1956), and whether most people implicitly believe that causal relationships are deterministic (Schulz & Sommerville, 2006; Frosch & Johnson-Laird, 2011). Our new model relies on FCMs, and is therefore most compatible with the position that people are causal determinists.

Although CBNs and FCMs capture different views about determinism, the similarities between these approaches should not be overlooked. Both approaches make use of probabilistic inference over structured representations, and both belong to a broader body of work that has used Bayes nets to account for multiple aspects of causal cognition (Holyoak & Cheng, 2011). For example, psychologists have used Bayes nets to model how people learn causal relationships (Gopnik et al., 2004), explain observed events (Halpern & Pearl, 2001), predict unobserved events (Gopnik et al., 2004), and interpret causal conditionals (Ali et al., 2011). The next section describes several models of counterfactual reasoning that belong to this same tradition, including the new model introduced in this paper.

Bayes net theories of counterfactual reasoning

This section describes several Bayes net models of counterfactual reasoning, including several previous models along with the new model that we have developed. Table 1 lists these models along with their distinguishing characteristics.

We will focus throughout on *retrospective* counterfactuals, or counterfactuals that relate to events that have already occurred, with premises that are contrary to fact. The bacon scenario in Figure 1 is one example in which certain events have taken place (e.g. the alarm has sounded) and we are interested in how these events might have turned out differently. Retrospective counterfactuals can be distinguished from *prospective* counterfactuals, which involve hypothetical future events. For example, a prospective counterfactual might state that if we cook bacon tomorrow, the smoke alarm will sound. Psychologists often focus on the differences between retrospective and prospective counterfactuals, but philosophers are more inclined to group them together as instances of hypothetical statements (Woodward, 2011). Each of the models in Table 1 can handle both retrospective and prospective counterfactuals, but we focus on retrospective counterfactuals because all of the models make the same predictions when applied to prospective counterfactuals.

Pearl's Structural Model

Pearl's theory of counterfactuals (Pearl, 2000, 2013), which we will call the *Structural Model* (SM), provides a starting point for all other models we consider. Pearl's approach relies on three basic assumptions. First, the SM relies on functional causal models, and is therefore committed to the idea that events which may appear stochastic would be revealed as deterministic if all of the relevant causal variables

Table 1
Distinguishing characteristics of the models we consider, including Pearl's Structural Model (SM), the Extended Structural Model (ESM), the Unattached Structural Model (USM), and the Minimal Networks Model (MNM).

Feature	SM	ESM	USM	MNM
Considers only counterfactual worlds that are minimally different from the actual world.	yes	no	no	yes
Represents noise/uncertainty in terms of hidden variables rather than inherently stochastic relationships.	yes	yes	no	no
Supports backtracking counterfactuals.	no	yes	yes	yes
Distinguishes between counterfactual intervention and observation.	no	yes	no	yes

were known. As described already, Figure 1 shows how a stochastic causal model for our bacon-cooking scenario can be converted into a functional model by adding hidden exogenous variables.

The SM's second assumption is that these exogenous variables retain their values in any counterfactual scenario one might imagine. For example, if the relationship between cooking bacon and the alarm activating is mediated by air currents, the SM assumes that these air currents are the same across the real and counterfactual worlds. This second assumption ensures that the SM considers only counterfactual worlds that are minimally different from the actual world, and can therefore be viewed as a version of the minimality assumption described earlier. Figure 2b illustrates how this minimality assumption can be captured using a *twin network*, in which every observable real-world variable has a corresponding counterfactual "twin".[1] Absent any counterfactual premise that differs from the real world, the values of the twin variables are identical to those of their counterparts, because they are deterministic functions of the same exogenous variables.

The SM's third assumption is that counterfactual premises are represented by imagining that an intervention has occurred in the counterfactual world. For example, when reasoning about a counterfactual world in which the smoke alarm does not activate, the SM assumes that the alarm was prevented from activating by a direct intervention. Such an intervention, like an idealized experiment, sets the value of the alarm variable to inactive while decoupling the alarm variable from its normal causes. Informally, under Pearl's account, the question "would B be true if S were true?" is equivalent to asking "would B be true if you had forced S to be true without influencing its causes or introducing any side effects?" The SM thus predicts that only the direct and indirect effects of a counterfactual premise will differ between the real and counterfactual worlds.

Given these assumptions and a specification of the causal relationships between variables, the SM makes precise predictions about which counterfactual inferences follow from a particular premise. The model has achieved a number of successes. It makes intuitive predictions about how people should distinguish between statements like "if Oswald didn't kill Kennedy, then someone else did" and statements like "if Oswald hadn't killed Kennedy, then someone else would have" (Adams, 1970; Pearl, 2013). The SM also shows how the effects of hypothetical experiments can be estimated from historical data (Shpitser & Pearl, 2009), and how one can identify the ways in which experimental results generalize to new contexts (Pearl & Bareinboim, 2011).

Despite these successes, the SM sometimes makes predictions that differ from human judgments, which limits its value as a psychological (as opposed to normative) theory. For example, unlike humans, the SM never makes backtracking inferences. Recall that backtracking occurs when a reasoner adjusts a cause in order to explain a counterfactual premise involving an effect. For example, a reasoner may decide that if the smoke alarm had not activated, then bacon would probably not have been cooked. This backtracking inference seems natural if the smoke detector is known to be reliable, and backtracking can be expected in other cases in which the relationship between cause and effect is highly reliable. For instance, if a cat falls into a pond and gets wet, it seems natural to infer that if the cat had not become wet, it would probably not have fallen into the pond.

A second limitation of the SM is that it always treats a counterfactual premise as an intervention, regardless of the specific character of that premise. In non-counterfactual causal reasoning tasks, people draw a strong distinction between observations and interventions (Hagmayer, Sloman, Lagnado, & Waldmann, 2007), and Sloman and Lagnado (2005) have shown that people are more prone to making

[1] Richardson and Robins (2013) discuss some limitations of twin networks and propose a different way of capturing counterfactual worlds using graphs. Shpitser and Pearl (2008) also propose a graphical representation that improves on twin networks. We will not discuss these approaches here because both make the same predictions as the twin-network approach for all of the applications we consider.

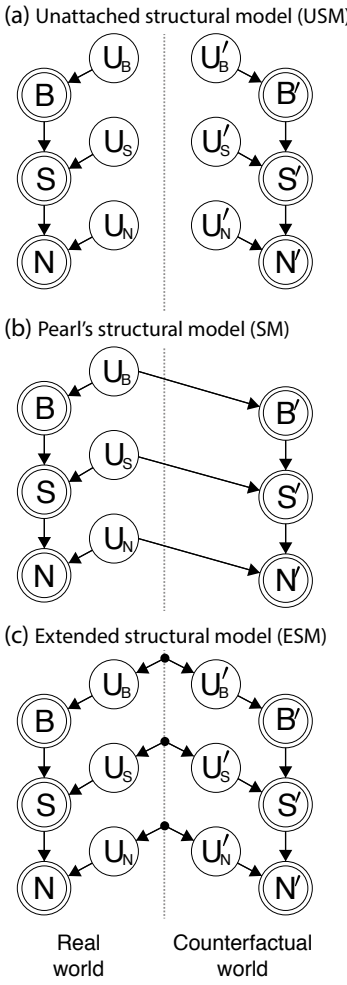


Figure 2. A comparison between three accounts of counterfactual reasoning using a causal chain scenario in which B causes S, which in turn causes N. The nodes B', S' and N' represent counterfactual values of B, S, and N. (a) An account that neglects the true state of the world when making counterfactual inferences, and treats the counterfactual world as if it were a completely new scenario. We refer to this account as the unattached structural model (USM). Exogenous variables are denoted with U_B, U_S and U_N, and nodes with double-edged boundaries are deterministic functions of their parents. (b) Pearl's Structural Model (SM) represented using a *twin network*. Information from the real world is used to make inferences about the exogenous variables, which in turn inform counterfactual inferences. (c) The Extended Structural Model (ESM), which extends the SM's twin network by allowing exogenous variables to take different values in the real and counterfactual worlds.

backtracking inferences when a premise is a counterfactual observation than when the premise is an intervention. For example, a counterfactual scenario in which the smoke alarm is simply observed to be inactive seems more likely to produce backtracking than a scenario in which the alarm is forced to be inactive.

The SM allows counterfactual observations to be captured in two different ways, but neither approach leads to backtracking. The first approach represents the observation of an event B by introducing a causal variable B_{obs} that is an effect of B. An observation of B is treated as an intervention that sets the value of B_{obs} and severs the link between B_{obs} and its parent. As a result, a counterfactual observation of B is uninformative about B's value in the counterfactual world, and therefore uninformative about the causes of B. The second, more conventional approach is to treat an observation of event B as information about the actual state of B. In this case, the SM treats a counterfactual observation like any other counterfactual premise, and the observed value of B still tells us nothing about B's causes in the counterfactual world.[2]

The limitations of the SM suggest that a successful psychological model of counterfactual reasoning should permit backtracking under some circumstances and should explain how humans distinguish between counterfactual observations and counterfactual interventions. One might argue that the SM is a normative model, and should not be expected to succeed as a psychological account. The next section, however, describes a simple extension of the SM that satisfies our psychological desiderata while inheriting many of the SM's appealing properties.

The Extended Structural Model

The SM assumes that all exogenous variables in the counterfactual world take values that match their real-world values. We now describe a new model that relaxes this minimality assumption and allows for the possibility that the exogenous variables take counterfactual values that differ from their real-world values. For example, suppose again that variable U_S in Figure 1 represents whether or not the prevailing air currents allow bacon smoke to reach the alarm. Even if U_S is true in the real world, our new model allows for the possibility that the air currents may be different in the counterfactual world. Our model builds on several key ideas behind the SM, including the idea that causal relationships are revealed as deterministic once all relevant variables have been taken into account. We therefore refer to our approach as the Extended Structural Model, or ESM for short.

[2] A third alternative is to set aside the twin network and interventions in the case of counterfactual observations, and instead treat them as non-counterfactual observations. We report the performance of this model in Appendix C.

The ESM includes a stability parameter s that captures the extent to which exogenous variables are expected to match across the real and counterfactual worlds. If stability is high (s close to 1), then the exogenous variables are very likely to match across these worlds. If stability is low (s close to 0), then the real and counterfactual exogenous variables are only weakly coupled. The role of stability can be quantified as follows:

$$P(U_i'|U_i) = s\delta(U_i) + (1-s)P_i(U_i'), \quad (1)$$

where U_i is an exogenous variable in the real world and U_i' is its counterfactual counterpart. Equation 1 can be interpreted as a probabilistic recipe for generating U_i'. With probability s, the value of U_i' is generated by simply copying the value of U_i, and with probability $(1-s)$, the value of U_i' is sampled from $P_i(\cdot)$, the prior distribution over U_i. The delta distribution $\delta(U_i)$ is a probability distribution that takes value 1 at U_i and value 0 at every other point.

If a causal model includes multiple exogenous variables, it is straightforward to allow different values of the stability parameter for each of these variables. Previous research has shown that some variables are more *mutable* than others—that is, more likely to be altered when reasoning about counterfactual scenarios (Kahneman & Miller, 1986). Using different stability values for different exogenous variables provides a way to capture this insight. Our experiments, however, deliberately use variables that are all of the same type, and we will therefore use a single value of the stability parameter for all exogenous variables. Formulating the model in this way minimizes the number of numerical parameters and therefore allows for a relatively strong test of our approach.

The predictions of the ESM can be captured using the extended twin network in Figure 2c. Unlike the standard twin network in Figure 2b, the extended network includes exogenous variables in the counterfactual world that may take values different from their corresponding variables in the real world. In general, the values assigned to the counterfactual exogenous variables will be influenced by the values of the real-world exogenous variables, by prior beliefs, and by the information conveyed by counterfactual premises.

Unlike the SM, the ESM can make backtracking inferences and can distinguish between counterfactual observations and counterfactual interventions. We illustrate by explaining how the extended twin network in Figure 2c is used to reason about observations and interventions. Figure 3 shows examples based on our bacon scenario. As before, we assume that B (bacon cooked), S (smoke alarm activates) and N (neighbors disturbed) are true in the real world, and that S' is false in the counterfactual world. Figure 3 shows how the extended network can be used to reason about a counterfactual observation of S. In this case we set B, S and N to true and S' to false, then compute the resulting probability distribution on B. As described in Appendix A, this computation can be carried out using any standard algorithm for inference in Bayesian networks. When $s = 0.5$, the model in Figure 3a predicts that $P(B' = t) = 0.487$. In other words, the model makes a backtracking inference and concludes that bacon was relatively unlikely to have been cooked in the counterfactual world.

Figure 3b shows how the extended twin network is used to reason about the scenario in which S' is fixed by a counterfactual intervention. As for the case just described, we set B, S and N to true. Consistent with the standard treatment of interventions in causal Bayes nets, we set S' to false and "mutilate" the graph by removing all incoming edges to S'. Adjusting the graph in this way captures the fact that the value of S' was determined by an intervention rather than by the parents of S' in the graph. Any standard algorithm for inference in Bayesian networks can then be used to compute the probability that $P(B' = t)$. When $s = 0.5$, the model in Figure 3b predicts that $P(B' = t) = 0.95$. As suggested by these examples, the ESM is compatible with the finding that people are more likely to backtrack given counterfactual observations than given counterfactual interventions (Sloman & Lagnado, 2005).

Two special cases of the ESM are important to consider. If the exogenous variables are maximally stable ($s = 1$), then the exogenous variables must match across these worlds, and the ESM reduces to the SM. Because the exogenous variables must match, the extended network in Figure 2c becomes equivalent to the standard twin network in Figure 2b. If stability takes its minimal possible value ($s = 0$), then U_i' does not depend on U_i, and these two variables are drawn independently from the distribution $P_i(\cdot)$. As a result, the extended network in Figure 2c becomes equivalent to the model shown in Figure 2a. We will refer to this special case as the USM, or "unattached structural model" because counterfactual variables are detached from their real-world twins. To our knowledge, the USM has not been proposed as a fully-general model of counterfactual reasoning, but it is closely related to the theory of counterfactual conditionals proposed by Over et al. (2007).

The special cases just considered suffer from different shortcomings. When $s = 1$, inferences about the counterfactual exogenous variables are not sensitive to the counterfactual premises. When $s = 0$, inferences about the counterfactual exogenous variables are not sensitive to the true state of the world. We propose that inferences about the counterfactual exogenous variables are sensitive to both the counterfactual premises and the true state of the world, and therefore expect that the judgments of most individuals will be consistent with stability values that lie between 0 and 1. All results reported later in the paper are based on setting $s = 0.53$, which is the value that provides the best account of

our experimental data.[3]

Equation 1 can be viewed as a weighted combination of the distribution on U_i' used by the SM and the distribution used by the USM. As a result, the ESM might initially seem equivalent to a weighted combination of the SM and the USM. This weighted combination would be consistent with the idea that some individuals always reason in accordance with the USM ($s = 0$) and others make inferences according to the SM's predictions ($s = 1$). A variation on the same idea is that each individual makes judgments consistent with the SM's predictions in some contexts and consistent with the USM's predictions in other contexts. Later, however, we show that these proposals can be distinguished empirically from the ESM. It turns out that a weighted combination at the level of the exogenous variables (Equation 1) does not lead to predictions that are weighted combinations at the level of the whole model: that is, the ESM's predictions are not weighted combinations of the predictions of the SM and the USM.

We presented Equation 1 as a probabilistic procedure for assigning values to the variables in the counterfactual world. The stability parameter s, however, can also be interpreted as a declarative characterization of the relationship between the real and counterfactual worlds. Suppose that these two worlds (along with many others) are generated by a branching process that produces a multiverse of possible worlds. The stability parameter s can be defined as a decreasing function of the time that has elapsed since the two worlds separated. Figure 2c shows two worlds, and the small black dots capture the states of the exogenous variables at the instant when the two worlds separated. The six arrows that issue from these dots are identical in length, and this length represents how long the two worlds have been evolving separately. As s approaches 1, the arrow length approaches zero, and each exogenous variable U' becomes effectively identical to U. As s approaches 0, the arrow length approaches infinity, and U' and U effectively become independent draws from the same prior. Full details are provided in Appendix A, which includes a derivation of Equation 1.

CBN models of counterfactual reasoning

As mentioned earlier, Pearl has argued that functional causal models (FCMs) provide a transparent account of counterfactual reasoning, and the two models presented so far both rely on FCMs. There are, however, alternative approaches to counterfactual reasoning that rely on stochastic causal Bayes nets rather than FCMs. Here we focus on a proposal due to Rips, who adapted Hiddleston's (2005) theory of counterfactuals to reconcile it with psychological data.

Rips's *Minimal Networks Model* (MNM) is founded on a minimality assumption, and proposes that people evaluate counterfactual statements by considering only counterfactual worlds that do not unnecessarily alter or "break" any causal relationships. Figure 4 gives an example of a causal system which describes four variables B, T, S, and N and the causal relationships among them. We can imagine this as an extension of our bacon-cooking scenario, so that B denotes cooking bacon, which is almost always true, T denotes burning toast, which is sometimes true, and the smoke alarm activating, S, is always true if either of those causes is present. For simplicity, we now assume that both S and N are deterministic effects of their parents in Figure 4: for example, the state of the neighbor variable N always matches the state of the alarm variable S. Suppose that none of the four events occurred in reality—all of the variables take the value f—and we are asked "if the smoke alarm had activated, would the bacon have been cooked?" The probability that the alarm activates in the absence of B and T is zero, so it is necessary to consider counterfactual worlds in which B or T is true. However, given that both B and T are sufficient causes of S, there is no requirement that they *both* be changed. Thus there are two minimal networks, shown in Figure 4(b). Because variable N has a cause (S) that is necessarily different in the counterfactual world, N is said to be "up for grabs" rather than a break, and can thus take any value that has a non-zero probability given that S is true.

How does the MNM map a set of minimal networks on to inferences about cooking bacon? Rips offers two approaches. The first is to count the number of minimal networks in which B is true, and divide that by the total number of minimal networks, yielding $P(B' = t) = 0.5$. The second approach is to weight networks by how probable they are: that is, by the joint probability of all of their variables given the counterfactual premise. In the current example, the second approach yields $P(B' = t) = 0.95$, because the base rate of B exceeds the base rate of T, and as a result the first minimal network in Figure 4 is more probable than the second. In practice, the MNM makes predictions using a mixture of these two approaches and random guessing, with the precise mixture of these strategies determined by two free parameters.

A noteworthy feature of the MNM is that non-minimal networks never contribute to inferences, no matter how probable they are. In Figure 4(b), for example, the bottom row is out of consideration, no matter how likely it is that events B and T are true in general. In contrast, the ESM does not make a minimality assumption, and allows for the possibility that B and T are both true in the counterfactual world. Our experiments that contrast the MNM and the ESM will explore this difference between the models, among other issues.

The original presentation of the MNM did not address the difference between counterfactual interventions and counterfactual observations. Rips and Edwards (2013), however, point out that the theory can be extended to accommodate this distinction, and we will evaluate the extended MNM ap-

[3] We also optimized the parameters used by all alternative models; see Appendix C for details.

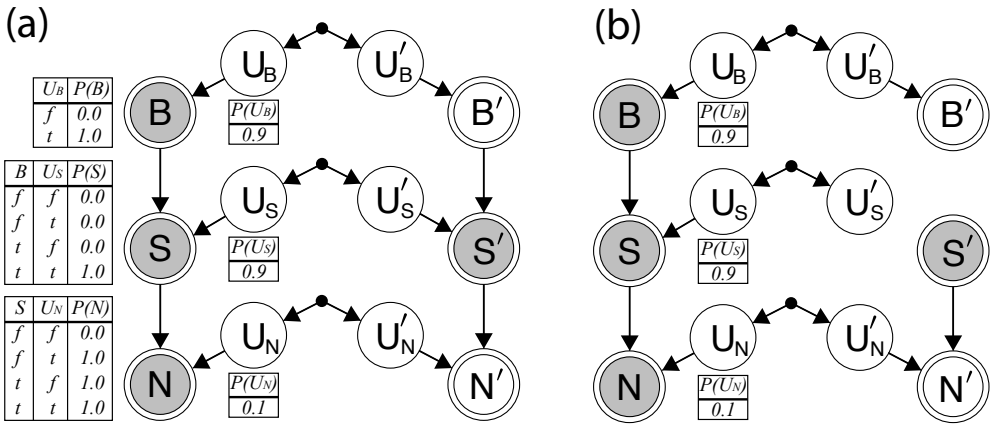
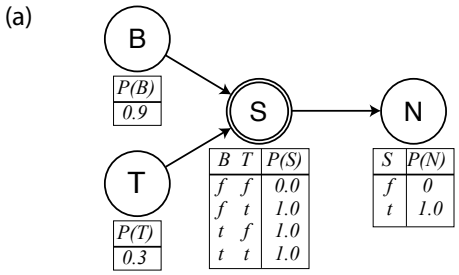


Figure 3. Differences between counterfactual observations and interventions under the ESM. Shaded nodes denote variables that have known values due to observation or intervention. (a) An example of a counterfactual observation when the true values of B, S, and N are known, and the counterfactual premise is that S' is observed to be different. (b) The same scenario under the counterfactual premise that an intervention has changed S'.



(b)

Scenario	*B*	*T*	*S*	*N*
Real-world	*f*	*f*	*f*	*f*
Counterfactual	*?*	*?*	*t*	*?*
Minimal 1	*t*	*f*	*t*	*t*
Minimal 2	*f*	*t*	*t*	*t*
Non-minimal	*t*	*t*	*t*	*t*

Figure 4. Minimality under the Minimal Networks Model. (a) A causal system in which a variable S has two deterministic, sufficient causes, along with a table giving the probabilities of effects given their causes. Each variable has two possible values: t, for *true*, and f, for *false*. (b) The real-world values of the variables, the counterfactual premise, two worlds with minimal breaks (breaks are denoted with gray backgrounds), and a world with non-minimal breaks.

proach that they briefly describe. To illustrate how interventions are handled, consider the scenario in Figure 4, and suppose that the counterfactual premise states that S had been activated by means of an intervention. We add a new cause of S to represent this intervention, and expand the conditional probability table for S to capture the idea that the intervention fixes the value of S. After expanding the network in this way, standard MNM theory can be used to make inferences about the other variables in the network.

As mentioned earlier, two key features of the MNM are that it relies on stochastic models and that it does not treat all counterfactual premises as interventions. Dehghani et al. (2012) have developed a second psychological model shares both features. According to Dehghani et al., backtracking is an incremental process that only occurs if the counterfactual premise is surprising in light of its causes. If some threshold level of surprise is crossed, then those causes are assigned new values such that the premise is no longer surprising.

If those causes' new values are themselves surprising, their causes are updated as well, with the process continuing until no relationship is too surprising, or no variables remain to be updated. Unlike Rips's MNM, Dehghani et al.'s model has several aspects that are not formally specified, such as the possibility that a learner's tolerance for surprise changes over time and the conditions under which multiple causes are simultaneously selected for updating. Given these ambiguities, and the broad similarities between Dehghani's and Rips's models, we will focus our attention on Rips's approach.

We have now introduced several different models of counterfactual reasoning that are summarized in Table 1. The first is Pearl's SM, which considers counterfactual worlds that are maximally similar to the actual world in the sense that they have matching exogenous variables. The second is the ESM, which balances the similarity of counterfactual worlds to the actual world against the prior probability of these counterfactual worlds. The ESM inherits the SM's approach to noise and variability, but differs from the SM by allowing backtracking inferences and distinguishing between counterfactual interventions and counterfactual observations. Along with the ESM, we described some related special cases and alternatives, such as the USM, which ignores real-world values of variables, and mixtures of the SM and USM. Finally, we described two existing models that allow causal relationships to be intrinsically stochastic, and that explain backtracking in a way that is related to Hiddleston's theory of counterfactuals. The next section describes a set of experiments that compares these models and assesses the specific commitments of the ESM.

Experiments

We conducted several experiments with the goal of testing the commitments of the ESM and distinguishing it from other models on the basis of the features listed in Table 1. All of our experiments shared the same basic form, and presented participants with information about causal structure, probabilistic information, information about the true state of the world and counterfactual premises. On the basis of that information, participants made judgments about counterfactual and non-counterfactual scenarios. One of our aims was to assess the models' ability to predict patterns of judgments by individual participants, so we conducted all of our experiments in a single batch: each participant saw every condition of every experiment. The experiments were presented in random order in a single session.

Experiment 1 sets the stage for subsequent experiments, and establishes that people expect that variables tend to be stable between the real and counterfactual worlds given the kinds of scenarios and cover stories that we present. In Experiment 2, we assess backtracking behavior given counterfactual observations and interventions, and compare the ESM's predictions to those of Pearl's SM. Experiment 3 explores whether people's counterfactual inferences are consistent with the ESM's commitment to deterministic relationships and exogenous hidden causes. Experiments 4 and 5 compare the ESM to the Minimal Networks Model as described by Rips (2010), and Experiment 6 compares the ESM to mixtures of different strategies. The counterfactual conditions in all experiments were conducted on a within-subjects basis, with each participant participating in every experiment in a randomly-determined order. Some experiments also included non-counterfactual baseline conditions, which were conducted in a separate group, with all baseline participants seeing every baseline condition in randomly-determined order.

For the experiments that follow, we report mean human judgments and predictions of each of the models we have discussed. Both the ESM and the MNM have numeric parameters. The ESM's predictions depend on a single parameter that specifies the stability of the exogenous variables. The MNM has two parameters in total: one determining the rate at which participants guess randomly, and a second that captures the extent to which they incorporate probabilistic information. For both models, we used a single parameter or parameter vector that minimized the discrepancy between model predictions and human responses across all experiments, rather than fitting parameters on an experiment-by-experiment basis. Best-fitting values of the parameters are provided in Appendix C. To make the core ideas of the MNM as clear as possible, the plots associated with each experiment show the MNM's predictions assuming that people do not guess randomly. Following the results for our individual experiments, we show overall model fits that are based on an optimized guessing parameter for the MNM.

Experiment 1: Preserving real-world values

A common idea behind theories of counterfactual reasoning, including our own, is that people try to preserve the true state of the world when reasoning about the consequences of a counterfactual premise. For example, suppose that two dice A and B are rolled and that both show the number 6. Imagine now that die A had not showed a 6. Intuition suggests that die B would probably still have showed a 6 in this counterfactual scenario. More generally, if some fact is true in the real world, that fact will probably remain true given any counterfactual premise that is unrelated to the fact.

Our first experiment used the simplest possible causal structure that allowed us to explore the preservation of real-world values. As shown in Figure 5a, the structure includes two variables that are not causally related to each other. Our cover story described a lab that was conducting research on hormones in mice, and described the variables as two hormones that were present in a particular mouse. This framing was designed to be compatible with a wide variety of different causal relationships between hormones, and to avoid

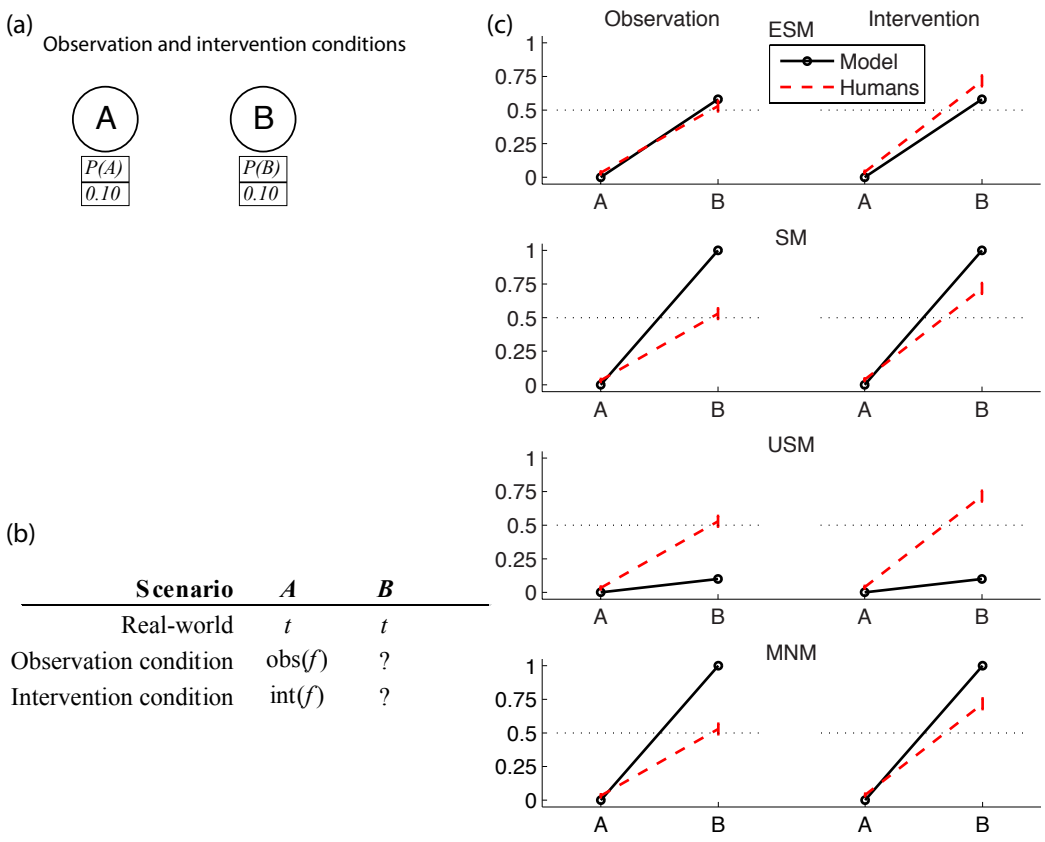


Figure 5. Materials and results for Experiment 1. (a) Causal graphical model representing the scenario in Experiment 1. In this system, hormones A and B are causally independent, so there is no edge between the nodes. The probability tables show the prior probability that each hormone is present. (b) Summary of the hormones' real-world states and their states in the two counterfactual scenarios, where *t* (true) and *f* (false) denote the presence or absence of a particular hormone, and '?' denotes a hormone with an unknown counterfactual value. (c) Participants' estimated probability judgments and model predictions. Error bars represent standard errors of the means in this and all subsequent figures.

strong a priori intuitions, which participants might have had about more commonplace scenarios. Figure 5b shows the real-world states of the hormones, as well as the premises in the two experimental conditions. In the *observation* condition, participants were asked whether *B* would be present if they had observed *A* to be absent. In the *intervention* condition, participants were asked whether *B* would be present if *A* had been absent as the result of an intervention.

According to the ESM people should be inclined to preserve the true state of hormone *B* given a counterfactual premise about hormone *A*, for both counterfactual interventions and counterfactual observations. The ESM makes this prediction because the exogenous variables in the counterfactual world are often expected to take values that match their real-world values. The model, however, also allows exogenous variables to differ across the real and counterfactual worlds, which means that inferences about hormone *B* are predicted to be below ceiling for both observations and interventions.

Several other outcomes are possible. For instance, the USM proposes that participants ignore the true state of the world when reasoning about counterfactual situations. The model therefore predicts that participants will not preserve the real-world state of hormone *B* when reasoning about a counterfactual observation of *A*. Alternatively, participants might make a minimality assumption and consider only counterfactual scenarios that differ in minimal respects from the actual world. If so, then participants should match the predictions of the SM and should preserve the true value of variable *B* with very high confidence.

Methods

Participants. We recruited participants through the psychology participant pool at Carnegie Mellon University and online, through Amazon's Mechanical Turk service. In-lab participants were randomly assigned to one of two

groups: a counterfactual group, with 55 participants who saw all counterfactual conditions for Experiments 1 through 6; and a baseline group with 39 participants who saw non-counterfactual versions of the conditions in Experiments 1, 2, and 6, for which non-counterfactual baseline data were relevant. Of the online participants, 56 were recruited for the counterfactual group, and 40 were recruited for the baseline group.

Materials. The observation and intervention conditions both used short stories along with a series of questions. The stories for this experiment and all subsequent experiments are reproduced in Appendix E. Both stories began with an introduction that mentioned a mouse that was the subject of the counterfactual questions, along with two hormones *A* and *B*, which could be present or absent in a given mouse. For both the observation and intervention conditions, the hormones *A* and *B* were said to be present in "a very small number" of mice. This verbal description was converted to the probabilities in Figure 5a by asking participants how many out of 100 mice "a very small number" corresponds to at the end of the study, and rounding the median (10) to the nearest multiple of 5. The same procedure was used to map verbal descriptions to probabilities in subsequent experiments. For each hormone, participants read that its presence or absence did not depend on the status of the other hormone, i.e., the two hormones were causally independent. Participants also read that both hormone *A* and hormone *B* were present in the current mouse.

After the story, each condition included a set of questions about each of the hormones that might be present in the current mouse, presented in random order. In the observation condition, participants were asked whether hormone *B* would be present if they had observed hormone *A* to be absent in that particular mouse. In the intervention condition, participants were asked whether hormone *B* would be absent if that mouse had been raised on a special kibble that specifically prevented the formation of hormone *A*. As a manipulation check, participants were also asked if hormone *A* would be absent if the mouse had been raised on special *A*-preventing kibble.

Each question included a yes/no question about whether the current hormone was present or not, followed by the question "How confident are you in your answer?" Confidence ratings were provided using a five-point scale where 1 was marked as "Not at all confident" and 5 was marked "Completely confident". We converted these ratings to subjective probabilities by treating "Not at all confident" responses as probabilities of 0.5, and linearly interpolating between 0.5 and either 0 or 1, where the direction of interpolation depended on the answer to the yes/no question. For example, an answer of "no" and a confidence of 3 was mapped to a subjective probability of 0.25, i.e., 0.5 minus 0.125 per point less than "Completely confident." An answer of "yes" and a confidence of 3 was mapped to a subjective probability of 0.75.

To understand how the counterfactual nature of our questions influenced people's judgments, we ran two additional conditions that did not involve counterfactuals but were otherwise identical to the observation and intervention conditions. In the *observation-baseline* condition, participants were told "you observe A to be absent" rather than being given non-counterfactual information followed by a counterfactual premise about *A*. Similarly, in the *intervention-baseline* condition, participants were simply told "you have raised Frank [the mouse] on special kibble", having been told only that the special kibble specifically prevents hormone *A* from being present.

Design and procedure. Every person in the counterfactual group participated in both counterfactual conditions (observation and intervention) in random order, as part of the larger set of experiments. Similarly, every person in the baseline group participated in the corresponding baseline conditions in random order, as part of the larger set of experiments. Questions could not be skipped, and participants could not return to a condition once they had submitted their answers.

Results and discussion

The counterfactual and baseline groups both received questions with answers that were unambiguously determined by the cover story, such as "If hormone *A* were present, would hormone *A* be present?" or "If you had observed hormone *A* to be present, would hormone *A* be present?" There were 9 such questions in the counterfactual group and 5 in the baseline group. We excluded participants who answered any of those questions incorrectly, leaving 30 in-lab participants in the counterfactual group and 30 in the baseline group. We applied the same exclusion criterion to online participants, leaving 34 remaining participants in the counterfactual group and 23 participants in the baseline group. Preliminary analyses revealed the same patterns of results across the two populations, and we combine them in our results.[4]

Figure 5c shows model predictions and participant judgments for both conditions. Mean human judgments for the hormones *A*, *B*, and *C* are displayed in red and repeated across rows, for comparison to the predictions of different models. Using derived subjective probabilities, we found that participants were more likely to say that hormone *B* was present in the observation condition ($M = .53$, $SD = .31$) than in the observation-baseline condition (Welch's t-test; $M = .41$, $SD = .26$; $t(115.0) = 2.21$, $p = .029$). Similarly, participants were more likely to to say that hormone *B* was present in the intervention condition ($M = .72$, $SD = .31$)

[4] Appendix C shows the accuracy of our model in predicting the in-lab and online groups when those groups are considered separately. Both accuracies are similar to the merged-group accuracy.

than in the intervention-baseline condition ($M = .46$, $SD = .28$; $t(114.4) = 4.69$, $p < .001$). These results indicate that participants tended to preserve the real-world states of the variables, as predicted by the ESM.

The remaining models in Figure 5 accounted less well for the data. The SM predicts that hormone *B* will take its real-world value in the counterfactual scenarios, leaving no room for uncertainty — nothing but *B*'s exogenous causes influence the counterfactual *B*, and these exogenous causes are unchanged in the counterfactual world. These stark predictions differ from the graded responses that humans offer, but a fairer evaluation of the SM might adopt the MNM's idea that participants are sometimes guessing randomly. We consider such a modification in Appendix C, and find that it does not greatly improve the overall performance of the SM even when the guessing rate is optimized based on our data.

At the opposite extreme, the USM predicts that people do not consider the real-world states of variables at all, and predicts that *B* will be counterfactually present in a very small number of mice, contrary to what people report. Like the SM, the MNM expects that people will expect hormone *B* to take its real-world value in the counterfactual scenario. The MNM, however, includes a parameter for the rate at which people guess randomly. By setting this parameter close to 1.0, which captures the idea that people are almost always guessing, the MNM can explain the current data. Experiments 2 through 5, however, used similar but more complex materials and yielded mean judgments that were far from chance, which suggests that participants were not guessing randomly in the current experiment.

Although our results are broadly consistent with the ESM, one aspect of our data is not predicted by this model. Participants tended to preserve the true values of hormone *B* more strongly in the intervention condition than in the observation condition (binomial sign test, $p = .001$). One possible explanation is that some participants postulated a common hidden cause that is responsible for the joint presence of *A* and *B*. Given a hidden cause of this kind, the ESM predicts lower judgments for the observation condition than the intervention condition.

In summary, Experiment 1 suggests that people tend to preserve real-word outcomes in counterfactual scenarios, but do not do so universally or with high confidence. This result is naturally captured by the ESM, but contrary to the predictions of the SM and the USM.

Experiment 2: Observations and interventions

Experiment 1 focused on a very simple scenario for which the ESM makes identical inferences about observation- and intervention-based counterfactuals. We designed Experiment 2 with the aims of confirming that people interpret observational and interventional counterfactuals differently, and establishing that the ESM (unlike the SM) makes predictions that align with human judgments in these cases.

The experiment used a chain of hormones with the same causal structure as the system in our running bacon example. As shown in Figure 6b, all hormones in the chain were said to be present in the real world. In the *observation* condition, participants were asked what values *A* and *C* would have taken if they had observed *B* to be absent. The *intervention* condition was similar, except the counterfactual premise stated that *B* was absent as a result of an intervention.

All of the theories under consideration predict that *C* would probably have been absent if *B* had been absent. The theories, however, make different predictions about whether and when people will backtrack and adjust the status of *A* given the counterfactual premise involving *B*. As described previously using our bacon example, the ESM predicts that participants are more likely to backtrack in the observation condition than the intervention condition. The MNM makes the same prediction if extended as previously described to distinguish between counterfactual observations and counterfactual interventions. In contrast, the SM does not provide a way to distinguish between counterfactual observations and counterfactual interventions.

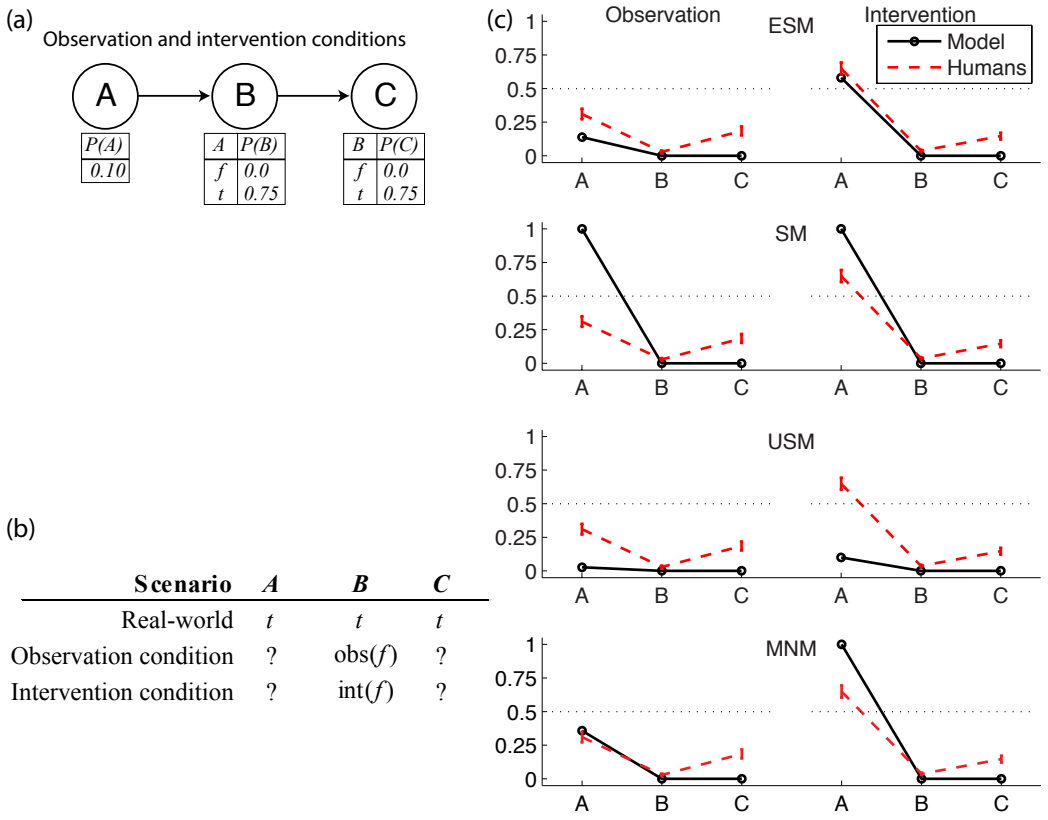
Our experiment is a conceptual replication of Sloman and Lagnado's (2005) Experiment 2. Sloman and Lagnado found that counterfactual observations were more likely to produce backtracking than counterfactual interventions, and we expected that our experiment would produce a similar result. The question of primary interest is therefore the extent to which the different models under consideration can account for this result.

Methods

Participants. Experiment 2 included the same set of participants as Experiment 1 and all other subsequent experiments.

Materials. Experiment 2 included observation and intervention conditions, consisting of stories and questions with the same basic structure as those in Experiment 1, but differing in the causal structures they described, the real-world states of the hormones, and the counterfactual premises as shown in Figure 6a-b. As in Experiment 1 and the experiments that follow, the probabilities in Figure 6a correspond to statements like "...in many of the mice...", where the numerical values were obtained by asking participants how many out of 100 mice "many" corresponded to, taking the median, and rounding to the nearest multiple of 5.

Both stories described a causal chain, shown in Figure 6a. The stories stated that in many of the mice hormone *A*'s presence causes *B* to be present, and in many of the mice hormone *B*'s presence causes *C* to be present. The stories also stated that hormone *A* is absent is almost all of the mice, that hormone *A*'s presence is the only cause of hormone *B*'s



(b)

Scenario	*A*	*B*	*C*
Real-world	*t*	*t*	*t*
Observation condition	?	obs(f)	?
Intervention condition	?	int(f)	?

Figure 6. Materials and results for Experiment 2. (a) Causal graphical models representing the scenarios in Experiment 2, where t and f denote the presence or absence of a particular hormone. (b) Summary of hormones' real-world states and their states in the two counterfactual scenarios (c) Participants' estimated probability judgments and model predictions.

presence, and that hormone *B*'s presence is the only cause of hormone *C*'s presence.

Participants in both conditions read that hormones *A*, *B*, and *C* were all present in the mouse, but the counterfactual premises varied by condition. In the observation condition, participants were asked whether each of the hormones would be present if they had observed hormone *B* to be absent in the current mouse. In the intervention condition, participants were asked whether hormones *A* and *C* would be present if the current mouse had been raised on a special kibble that specifically prevented the formation of hormone *B*.

As in Experiment 1, we compared our participants' judgments to judgments from non-counterfactual baseline conditions, *observation-baseline* and *intervention-baseline*, that were otherwise identical to the observation and intervention conditions. In the observation-baseline condition, participants were told "you observe B to be absent" rather than being given non-counterfactual information followed by a counterfactual premise about *B*. In the intervention-baseline condition, participants were simply told "you have raised Nellie [the mouse] on special kibble", having been told only that the special kibble specifically prevents hormone *B* from being present.

Design and procedure. The procedures for Experiment 2 were identical to those used in Experiment 1 and all other subsequent experiments.

Results and discussion

Mean responses and model predictions are shown in Figure 6c. The Bayes net in Figure 6a is not a functional causal model, and the predictions of the SM and ESM are therefore based on the corresponding functional model in Appendix B. As predicted by the ESM, participants preserved the real-world value of the counterfactual premise's cause (hormone *A*) more often in the intervention condition than the observation condition (binomial sign test, $p < .001$). In the intervention condition, participants were more likely to preserve the real-world value of hormone *A* than their counterparts in the baseline condition ($t(114.1) = 4.39$, $p < .001$). For the observation condition, there was no significant difference ($t(110.2) = .53$, $p = .60$).

The ESM successfully predicts that people will make

backtracking inferences when asked to reason about counterfactual observations, but will do so more weakly when the counterfactual premise is an intervention. In contrast, the SM predicts an absence of backtracking — meaning that A will be present — in both conditions, and the USM predicts backtracking in both conditions. The MNM distinguishes between the two kinds of counterfactuals by positing a new cause in the intervention case (Rips & Edwards, 2013), which makes any counterfactual change to the causes of the premise non-minimal. As a result, the MNM predicts no backtracking in the intervention case, but otherwise requires that a break occurs to explain the counterfactual value of B. Given that the available probabilistic information favors a break at A, the MNM predicts high levels of backtracking in the observation condition.

Experiment 2 provides evidence that people make backtracking inferences for both observational and interventional counterfactuals, expanding on the results of Sloman and Lagnado's (2005) Experiment 2. Whereas Sloman and Lagnado included non-causal control conditions, we included baseline conditions that were causal but not counterfactual, allowing us to focus narrowly on the question of how counterfactual inferences differ from their non-counterfactual counterparts.

Our baseline conditions also reveal that our data are inconsistent with a tentative proposal offered by Meder, Hagmayer, and Waldmann (2009). Among other issues, these authors explored whether people distinguish between counterfactual interventions and *hypothetical* interventions such as the intervention in our baseline condition. A critical difference between the two is that the state of the intervened-on variable is known in the case of a counterfactual intervention, but unknown in the case of a hypothetical intervention. Meder et al. reported no significant differences between inferences about counterfactual and hypothetical interventions, and raised the possibility that people may fail in general to distinguish between these kinds of interventions. This conclusion challenges both the SM and the ESM, which both treat counterfactual interventions and hypothetical interventions rather differently. Our data, however, suggest that people can distinguish between counterfactual and hypothetical interventions. As Meder et al. acknowledge, it is possible that their null result reflects the complexity of their task, which involved four variables and required participants to infer conditional probabilities from a set of 50 to 60 training events.

The most important finding from Experiment 2 is that people treat counterfactual observations and counterfactual interventions differently, which supports the fourth distinctive feature of the Extended Structural Model in Table 1. The specific pattern of results matches the predictions of the ESM, which posits a trade-off between the prior probability of a counterfactual world — as reflected in its consistency with base rates and known causal relationships — and the consistency of the counterfactual world with the real world. We now turn to another characteristic in Table 1 that distinguishes the ESM and the SM from the other accounts in the table. Both the ESM and the SM propose that people represent apparently stochastic causal relationships in terms of hidden exogenous variables, and that people prefer counterfactual scenarios in which those hidden variables take their real-world states.

Experiment 3: Stochastic versus deterministic relationships

A central tenet of the SM and the ESM is that any given causal relationship is deterministic at some level, and that people represent apparently stochastic relationships in terms of hidden, exogenous mediators. For example, people may explain the relationship between cooking bacon and alarm activation by invoking a mediating variable such as the direction of air currents. If people automatically reason in this way, then their inferences should be the same regardless of whether or not the mediating variable is explicitly described. Our third experiment tests this prediction and explores whether inferences about an apparently stochastic model are similar to inferences about a corresponding functional causal model.

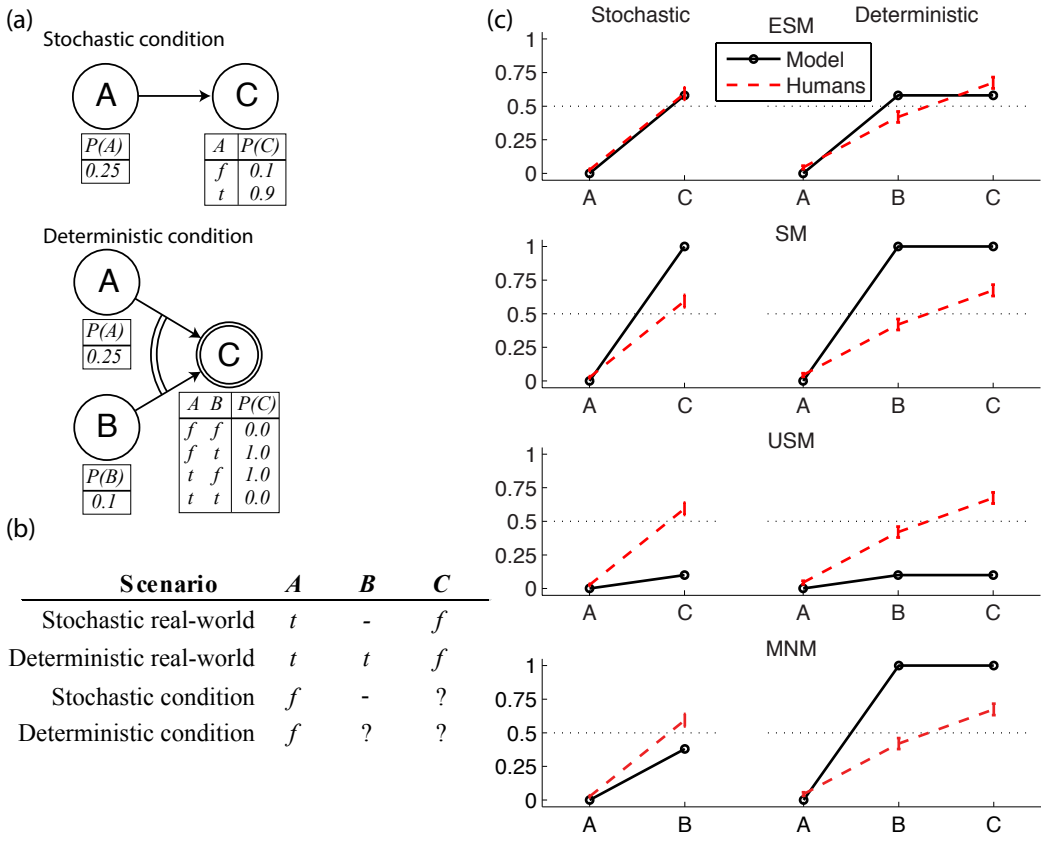
Experiment 3 asked people to reason about two structures shown in Figure 7a. The first structure specifies a stochastic relationship between hormones A and C, and the second specifies a deterministic relationship mediated by an additional hormone B. The base rate of B has been chosen so that the two structures capture the same probability distribution $P(C|A)$. The ESM and the SM predict that the stochastic structure is represented in a way that is equivalent to the deterministic structure, and therefore predicts that the two structures will lead to similar patterns of inference. In contrast, the MNM predicts that the two structures will be treated differently.

Methods

Participants. Experiment 3 included the same set of participants as Experiment 1 and all other subsequent experiments.

Materials. Participants saw *stochastic* and *deterministic* conditions, consisting of stories and questions with the same basic structure as those in Experiment 2, but differing in the causal structures they described, the real-world states of the hormones, and the counterfactual premises.

The two stories describe causal systems in which hormone A causes the presence or absence of hormone C, as shown in Figure 7a. In the stochastic condition, participants read that in a very small number of mice, hormone A's presence causes hormone C to be absent, and hormone A's absence causes hormone C to be present, while in the remaining mice



(b)

Scenario	*A*	*B*	*C*
Stochastic real-world	*t*	-	*f*
Deterministic real-world	*t*	*t*	*f*
Stochastic condition	*f*	-	?
Deterministic condition	*f*	?	?

Figure 7. Materials and results for Experiment 3. (a) Causal graphical models representing the scenarios in Experiment 3, where *t* and *f* denote the presence or absence of a particular hormone. The arcs between the arrows in the deterministic condition signify that the relationship between *A*, *B*, and *C* is not simply generative or preventative, but instead *B* flips the relationship between *A* and *C* from generative to preventative. Dashed edges indicate probabilistic relationships, and solid edges indicate deterministic relationships. (b) Summary of hormones' real-world states and their states in the two counterfactual scenarios. (c) Participants' estimated probability judgments and model predictions.

the presence of hormone *A* causes hormone *C* to be present and the absence of hormone *A* causes hormone *C* to be absent. In the materials given to participants and included in Appendix E, the effect variable in the stochastic condition was called *B* rather than *C*, but this variable has been relabeled in Figure 7a to emphasize that it plays the same role as the variable *C* in the deterministic condition. In the deterministic condition, the underlying relationship between *A* and *C* was similar, but now explicitly dependent upon a hormone *B* which was said to be present in a very small number of mice. If *B* is present then *C* is present in the absence of *A* and vice versa, and if *B* is absent then hormone *C*'s state matches that of hormone *A*. The conditional probability table for the deterministic condition in Figure 7a reflects this relationship: if *B* is present, *C* will be present only if *A* is absent, and if *B* is absent, *C* will be present only if *A* is present.

Figure 7b shows the real world states of the hormones along with the counterfactual premises. Participants in both conditions read that hormone *A* was present and hormone *C* was absent in the mouse, and read in the deterministic condition that hormone *B* was present. All participants were asked whether each of the hormones that had been identified would be present if hormone *A* were absent. Unlike Experiments 1 and 2, which used counterfactual observations and counterfactual interventions, Experiment 3 used a generic counterfactual premise which asked participants to think about "If A had been absent...' For the purpose of generating model predictions, we assume that generic premises of this kind are interpreted as counterfactual observations. The general discussion, however, considers some of the ways in which the interpretation of generic counterfactuals may be shaped by context.

Design and procedure. The procedures for Experiment 3 were identical to those used in Experiment 1 and all other subsequent experiments.

Results and discussion

Mean responses and model predictions are shown in Figure 7c. The results of primary interest are the inferences about variable *C* in both conditions. The ESM and SM make identical inferences in both conditions because they represent the stochastic structure in a way that is identical to the deterministic structure. According to the SM, *C* should reliably be present when *A* is counterfactually absent, because the value of the mediating variable *B* is preserved in the counterfactual world. The ESM predicts that there should be an intermediate level of confidence that *C* is present in the absence of *A*, because there is a tension between preserving the true value of the mediator and choosing a counterfactual world that is likely.

The MNM does not represent the stochastic structure in a way that matches the deterministic structure, and as a result makes different inferences about variable *C* in the two conditions. In the stochastic condition, the model relies on base rates and predicts that *C* is absent when *A* is counterfactually absent regardless of the real-world values of *A* and *C*. In the deterministic condition, the model infers that *B* is present in the real world, and carries the value of *B* over to the counterfactual world, which implies that *C* will be present.

The human responses in Figure 7c are more consistent with the ESM than the other models under consideration. Consistent with the ESM's predictions, and contrary to those of the MNM, there was no significant difference between judgments about hormone *C* in the two conditions (binomial sign test, $p = .36$). The average subjective probabilities that hormone *C* was present in the counterfactual world were .59 and .67 for the stochastic and deterministic conditions, respectively. The similar judgments for hormone *C* across the two conditions indicates that participants treated both conditions as involving a hidden mediator, rather than interpreting the stochastic condition as involving inherently stochastic relationships.

Although the ESM successfully predicts that inferences about *C* will be similar across the two conditions, none of the models explains why participants gave higher judgments for *C* than *B* in the deterministic condition (binomial sign test, $p < .001$). One might expect that inferences about *B* would be similar to inferences about variable *B* in Experiment 1, and it is therefore somewhat surprising that the mean judgment for *B* falls below 0.5 in the deterministic condition. The results for variable *B* may therefore indicate that a small number of participants found the causal relationships in the deterministic condition difficult to understand.

The ESM does not account for all aspects of our data, but Experiments 1 through 3 nevertheless support four basic commitments of the model that are summarized in Table 1. First, people appear willing to consider counterfactual scenarios that are more than minimally different from the actual world. Second, people's judgments about apparently stochastic relationships are consistent with inferences to hidden exogenous variables. Third, people reliably make backtracking inferences in some situations. Fourth, people interpret counterfactual observations and counterfactual interventions differently.

The results of our first three experiments raise some challenges for the Minimal Networks Model, but none of these challenges is conclusive. In particular, the MNM accounts fairly well for these experiments if its guessing parameter is set to a very high value. We now turn to a more direct comparison between the ESM and the MNM. The core assumption of the MNM is that people consider only minimal alterations of the real world, and Experiments 4 and 5 explore whether people's counterfactual inferences are consistent with this assumption. These experiments will allow us to distinguish between the ESM and the MNM regardless of whether the MNM's guessing parameter is small or large.

Experiment 4: Base rates and non-minimal breaks

We introduced the MNM using the example in Figure 4, in which cooking bacon and burning toast can both cause the alarm to sound. In reality, bacon was not cooked, toast was not burned, and the alarm did not sound. Suppose, however, that the alarm had sounded. According to the CPD for the alarm variable, the sounding of the alarm is possible only in worlds where either bacon was cooked, or toast was burned, or both. As shown in Figure 4, the MNM allows only minimal departures from the real world, and focuses on the two counterfactual worlds in which only one possible cause is present. The world in which bacon was cooked and toast was burned is non-minimal, because just one of these departures from the real world would be enough to explain the counterfactual premise (i.e. that the alarm sounded). Intuitively, however, if bacon-cooking and toast-burning are both common, it seems plausible that both might have occurred in the counterfactual world.

Experiments 4 and 5 use variants of the example just described to explore whether people consider non-minimal counterfactual worlds. Experiment 4 aims to distinguish between the ESM and the MNM by pitting the MNM's concept of minimality against probabilistic information. The causal structure indicates that hormones *A* and *B* can each separately cause hormone *C* to be present, (Figure 8a), and in the real world, hormones *A*, *B* and *C* are all absent in a particular mouse. If hormone *C* were present, what would be true of hormones *A* and *B*? Absent any additional information, both the ESM and the MNM predict that either *A* or *B* will be present, but probably not both. These models diverge, however, if *A* and *B* are almost always present in other mice. The MNM continues to predict that at most one sufficient cause will be present, because non-minimal states are excluded from consideration. In contrast, the ESM balances the preservation of real-world values against the probability

of different configurations, and predicts that hormones *A* and *B* are both more likely to be present than absent.

Methods

Participants. Experiment 4 included the same set of participants as Experiment 1 and all other subsequent experiments.

Materials. Experiment 4's materials consisted of one story with the same basic structure as those in Experiment 3, but differing in the causal structure it described, the real-world states of the hormones, and the counterfactual premise. The story described a causal system in which hormones *A* and *B* cause the presence or absence of hormone *C*, which in turn causes *D*, as shown in Figure 8a. Participants read that hormone *A*'s presence causes hormone *C* to be present in all mice, hormone *B*'s presence causes hormone *C* to be present in all mice, *C*'s presence causes hormone *D* to be present in all mice, and that hormones *A* and *B* are present in almost all mice. Participants read that all four hormones were absent in the current mouse, and were asked whether each of the hormones that had been identified would be present if hormone *C* were present.

We suspected that asking questions about the causes but not the effects of the counterfactual premise might make participants more prone to backtracking inferences, an issue we address in the general discussion. To mitigate this concern, we included an additional hormone *D*, which was an effect of *C*, and asked about it as well.

Design and procedure. The procedures for Experiment 4 were identical to those used in Experiment 1 and all other subsequent experiments.

Results and discussion

Figure 8c shows that participants reported that *A* and *B* would be present more often than not, for both *A* ($t(63) = 4.90$, $p < .001$) and *B* ($t(63) = 5.60$, $p < .001$). These results are consistent with the predictions of the ESM, which takes into account the high base rates of hormones *A* and *B* and permits both to be present in the counterfactual world. In contrast, the MNM only permits one of the two hormones to be present — changing both would be non-minimal — and thus predicts that *A* and *B* each have a .5 probability of being present.

Figure 8c also shows that the SM and the USM both fail to account for participants' inferences. The SM predicts that *A* and *B* will both take their real-world values, while the USM ignores the true state of the world and, unlike humans, simply matches A' and B' to their base rates.

The main conclusion from Experiment 4 is that probabilistic information can influence people's inferences about backtracking counterfactuals, even when minimality suggests that this information should be ignored. In this first contrast between the MNM and the ESM, we used a simple, symmetric causal scenario, and elicited a robust but relatively subtle effect. The next experiment uses a slightly more complex scenario, which allows us to explore a case in which the ESM and the Minimal Networks Model make starkly different predictions.

Experiment 5: Stochastic relationships and non-minimal breaks

Experiment 5 adapted the structure used in Experiment 4 as shown in Figure 9. As before, hormone *A* is a deterministic cause of *C*, but now hormone *B* is a stochastic cause — its presence *almost* always causes *C* to be present. In the real world, hormones *A*, *B* and *C* are all observed to be present. Consider now a counterfactual premise that states that *C* is absent. Under the MNM, the deterministic cause *A* must be absent in order to explain the counterfactual premise, and the noisy cause *B* must be present, as its absence would constitute a non-minimal break. In contrast, the ESM predicts that both causes are likely to be absent, because *C*'s absence would be highly unlikely even in the presence of *B* alone.

Methods

Participants. Experiment 5 included the same set of participants as Experiment 1 and all other subsequent experiments.

Materials. Experiment 5's materials consisted of one story with the same basic structure as that in Experiment 4. The story described a causal system in which hormones *A* and *B* cause the presence or absence of hormone *C*, which in turn causes *D*, as shown in Figure 9a. Participants read that hormone *A*'s presence causes hormone *C* to be present in all mice, hormone *B*'s presence causes hormone *C* to be present in almost all mice, and *C*'s presence causes hormone *D* to be present in all mice. Participants read that all four hormones were present in the current mouse, and were asked whether each of the hormones that had been identified would be present if hormone *C* were absent.

Design and procedure. The procedures for Experiment 5 were identical to those used in Experiment 1 and all other subsequent experiments.

Results and discussion

Model predictions and human responses are shown in Figure 9c. The MNM and the ESM both predict that *A* must be absent, because *C* cannot be absent in the presence of *A*. For hormone *B*, however, the models make very different predictions. The MNM predicts that *B* must be present, because changing *B* from its real-world value constitutes a superfluous change from the real world — a non-minimal break. The ESM predicts that *B* is relatively unlikely to be present, because it permits *B* to differ from its real-world value and it

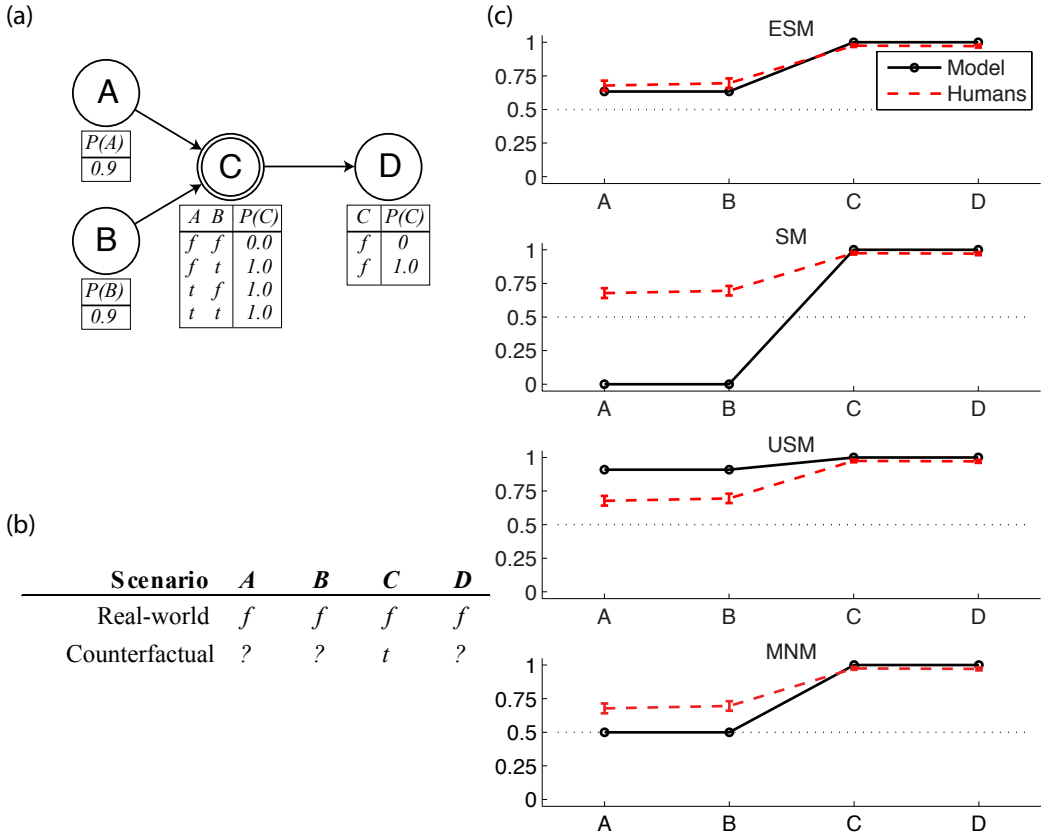


Figure 8. Materials and results for Experiment 4. (a) Causal graphical models representing the scenarios in Experiment 4, where *t* and *f* denote the presence or absence of a particular hormone. (b) Summary of hormones' real-world states and their states in the counterfactual scenario. (c) Participants' estimated probability judgments and model predictions.

is implausible, although possible, that C is absent when B is present.

We cannot evaluate the models by comparing average subjective probabilities for A and B to zero and one respectively, because variability in individuals' confidences could lead us to falsely conclude that the MNM's predictions are inaccurate. We can, however, compare judgments for hormones A and B to distinguish between the two models: the ESM predicts that judgments about B should be further from 1.0 than judgments of A are from zero. In contrast, the MNM predicts that A judgments should deviate from the floor at the same rate that B judgments deviate from ceiling, because any such deviations are due to cases where participants guess randomly.

Consistent with the ESM's predictions, participants' judgments for B were significantly further from ceiling than A judgments were from floor, as revealed by a hypothesis test comparing the sum of the two judgments to 1.0 ($t(63.0) = 4.57$, $p < .001$). This result indicates that people are willing to consider counterfactual worlds that are more than minimally different from the real world, especially when there is reason to believe that such counterfactual worlds are probable.

One notable difference between the predictions of all models and our participants' judgments is that participants did not confidently predict that hormone D, the effect of the counterfactual premise, would be absent. One possible explanation is that the cover story did not exclude the possibility that hormone D had other causes — if participants expect that an additional, unobserved cause of D may be present, then all of the models assign a non-zero probability that D will be present in both of the counterfactual scenarios.

Taken together, the results of Experiments 4 and 5 indicate that the ESM gives a better account of counterfactual inferences than the Minimal Networks Model. Combining these results with the results of Experiments 1 through 3 suggests that the ESM performs better than the three other models represented in Table 1—the SM, the USM, and the MNM. It remains possible, however, that some combination of these models might account for people's inferences. For example, perhaps individuals have access to multiple strategies for reasoning about counterfactual scenarios, and apply the SM in

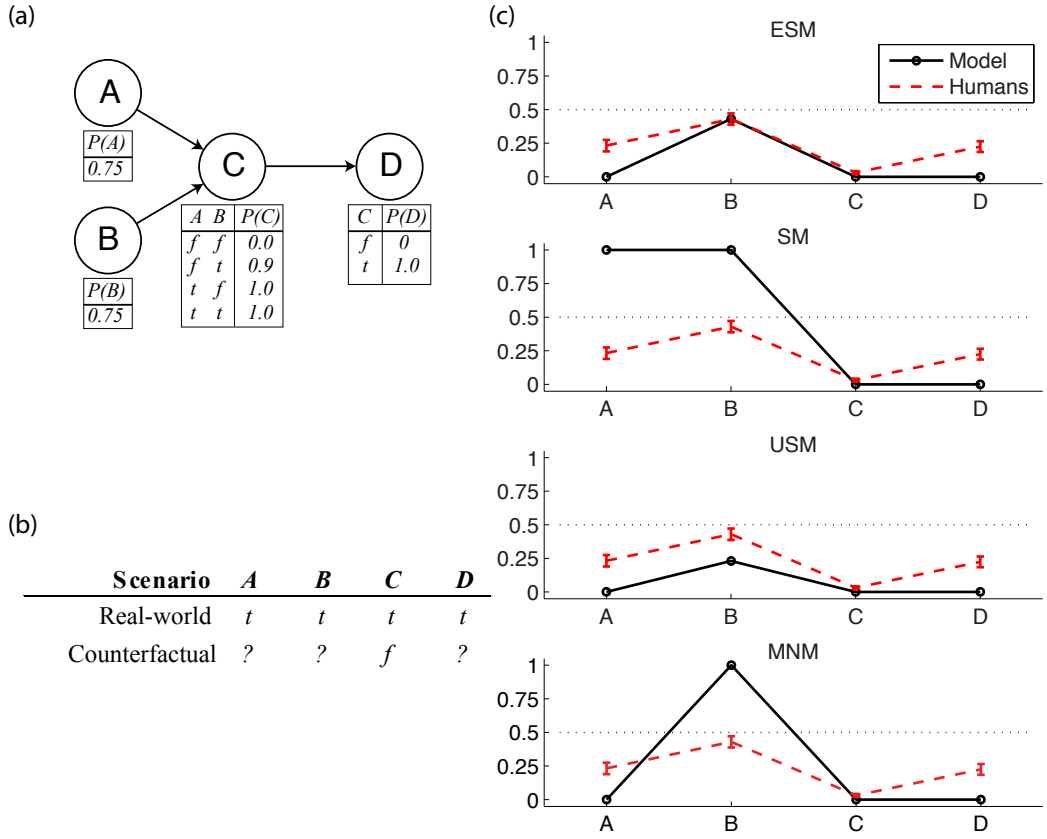


Figure 9. Materials and results for Experiment 5. (a) Causal graphical models representing the scenarios in Experiment 5, where *t* and *f* denote the presence or absence of a particular hormone. (b) Summary of hormones' real-world states and their states in the counterfactual scenario. (c) Participants' estimated probability judgments and model predictions.

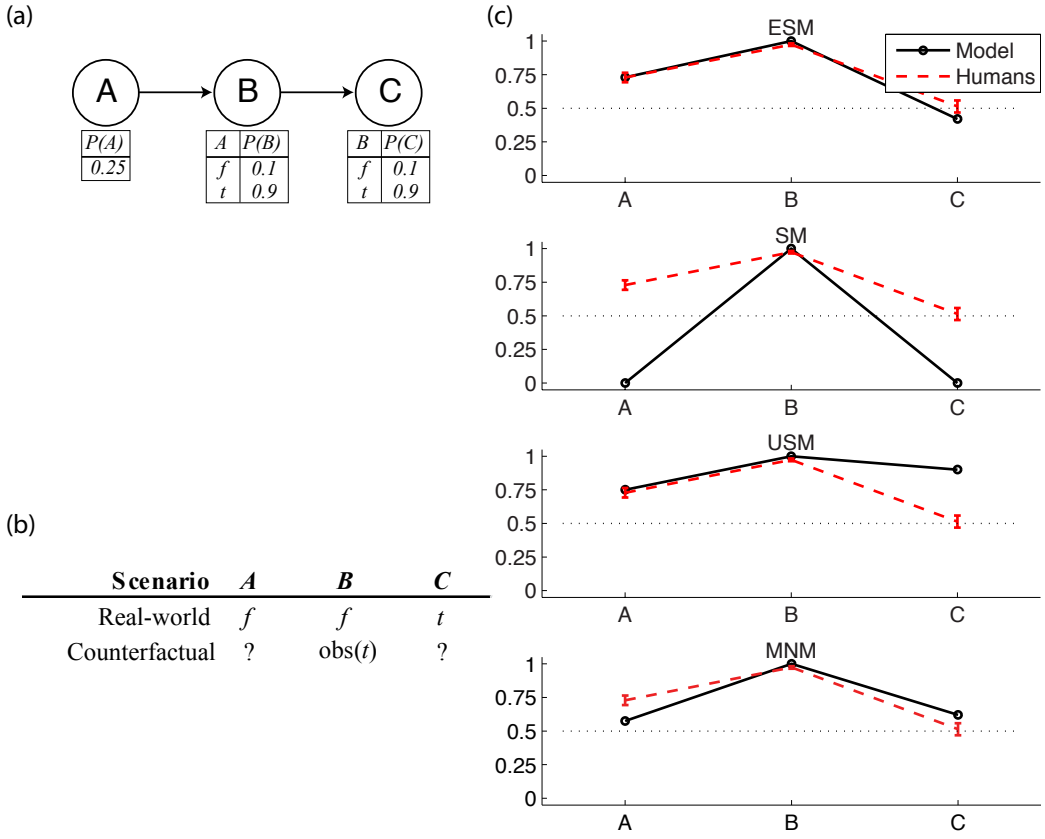
some contexts and the USM model in others. Our final experiment focuses on a scenario that distinguishes the ESM from any mixed strategy that relies on combinations of the SM, the USM, and the MNM.

Experiment 6: ESM versus mixed strategies

As described earlier, the SM and the USM are special cases of the ESM, and it is therefore natural to ask whether the ESM is empirically distinguishable from a weighted combination of the SM and the USM. From a psychological perspective, there are multiple ways in which the SM and the USM could be combined. One possibility is that each person consistently makes inferences consistent with one of the models, given that we have been looking at judgments averaged across individuals. Another possibility is that people might occasionally ignore the counterfactual nature of a scenario or query, and make judgments consistent with the USM. A third possibility is that a single individual might be uncertain about how to interpret a counterfactual query, and might therefore make a single judgment that averages over different models. In all of these cases, mean judgments over a sample will reflect a weighted average of the predictions of the SM and the USM.

The causal chain in Figure 10a allows us to distinguish between the ESM and weighted combinations of the SM, the USM and the MNM. In almost all mice, the value of B matches the value of A—in other words, the presence of A causes B to be present, and the absence of A causes B to be absent. In the remaining mice, the value of B is the opposite of the value of A. The relationship between B and C is similar: in almost all mice, the value of C matches the value of B, but in the remaining mice, the value of C is the opposite of the value of B.

In the real world, A and B are known to be absent and C is known to be present. This information suggests that the mouse in question is a mouse for which B matches A. Given the counterfactual premise that B is present, the ESM is inclined to preserve the fact that A matches B, and therefore predicts that A is probably present. The real-world values also suggest that the mouse is one of the unusual cases in which C does not match B. The ESM infers that C may continue to be different from B in the counterfactual world, and



(b)

Scenario	*A*	*B*	*C*
Real-world	*f*	*f*	*t*
Counterfactual	?	obs(*t*)	?

Figure 10. Materials and results for Experiment 6. (a) Causal graphical models representing the scenarios in Experiment 6, where *t* and *f* denote the presence or absence of a particular hormone. (b) Summary of hormones' real-world states and their states in the counterfactual scenario. (c) Participants' estimated probability judgments and model predictions.

therefore predicts that *C* is less likely to be present than *A*.

None of the remaining models generates higher predictions for *A* than for *C*. The SM predicts that *A* and *C* should both be absent. The prediction about *A* follows from treating the counterfactual premise as an intervention on *B*, which implies that *A*'s value in the counterfactual world will match its real-world value. The prediction about *C* follows from the inference that the mouse of interest is an unusual case in which *C* is different from *B*.

The USM and the MNM predict that *C* is more likely to be present than *A*. The USM makes this inference because the base rates of the three hormones increase along the chain—in particular, *B* and *C* are more likely *a priori* to be present than *A*. The MNM yields the same qualitative prediction as the USM, because all possible counterfactual worlds involve minimal breaks with respect to the real world.

Given these separate predictions, any weighted combination of the MNM, the SM, and the USM predicts that hormone *C* is at least as likely to be present as hormone *A*. Experiment 6 is therefore a strong test of the ESM's prediction that *C* is less likely to be present than *A*.

Methods

Participants. Experiment 6 included the same set of participants as Experiment 1 and all other subsequent experiments.

Materials. Experiment 6's materials consisted of one story with the same basic structure as that in Experiment 2, but differing in the causal structure it described, the real-world states of the hormones, and the counterfactual premise. Experiment 6 involved a slightly more detailed causal system than other experiments and may have seemed more complex to participants, so we included a non-counterfactual baseline condition — parallel to that used for Experiments 1 and 2 — to determine whether participants' non-counterfactual inferences were consistent with the probabilities and causal relationships described in the story.

The story described a causal system in which the state of hormone *A* causes the presence or absence of hormone *B*, which in turn causes the presence or absence of hormone *C*. Participants read that in almost all mice, if *A* is present, this causes *B* to be present, and if *A* is absent, this causes *B* to be absent. In the remaining mice, if *A* is present, this causes *B*

to be absent, and if A is absent, this causes B to be present.

The same relationship was described between B and C. In almost all mice, if B is present, this causes C to be present, and if B is absent, this causes C to be absent. In the remaining mice, if B is present, this causes C to be absent, and if B is absent, this causes C to be present.

Participants read that hormones A and B were present in the current mouse but C was absent, and were asked whether each of the hormones that had been identified would be present if hormone B were observed to be present.

Design and procedure. The procedures for Experiment 6 were identical to those used in Experiment 1 and all other subsequent experiments.

Results and discussion

Our results in the baseline condition, shown in Appendix D, indicate that participants understood the probabilities and causal relationships described in the cover story. Average responses to the counterfactual questions are shown in Figure 10c. Consistent with the ESM, people infer that hormone A is more likely to be present than hormone C (binomial sign test $p = .0039$). As described earlier, this result is inconsistent with any weighted combination of the SM, the USM, and the MNM. We can therefore conclude that the ESM provides a better account of people's counterfactual inferences than any mixed strategy involving the alternative models considered in this paper.

Overall model comparison

Previous sections described six individual experiments, and we now compare the overall performance of the models that we have considered. Figure 11 provides a quantitative summary of the performance of each model for each experiment. The measure used is sum squared error, and lower error rates indicate better performance. The mixture model plotted is a weighted combination of the ESM and the USM that relies on a single free parameter to specify the relative weights of the two models.

Figure 11 shows that the ESM had the lowest prediction errors for Experiments 1, 3, 4, and 6. In Experiments 2 and 5, the ESM yielded marginally higher prediction errors than the MNM and the mixture model, respectively. Overall, however, the ESM made the most accurate predictions across the six experiments.

We can also compare model performance by looking at the correlation between model predictions and mean human judgments. For the purpose of this comparison, we excluded all manipulation-check items, because responses to these items are not revealing about the differences between models. Figure 12 shows that the ESM's predictions correspond closely to human judgments, and produce the highest correlation with human judgments.

As mentioned previously, we fit each model by selecting a single parameter or parameter vector that minimizes its sum squared error across all judgments. The SM and the USM have no free parameters, the ESM and the mixture model include one free parameter each, and the MNM includes two free parameters. It is important to ask whether the performance of the ESM is highly sensitive to the values of its stability parameter. We addressed this question by examining fits of the ESM to human judgments as a function of the stability parameter. For comparison, we also plotted the sensitivity of the mixture model (the second most accurate model) to the weight it assigns to the USM versus the SM. The gray region in Figure 13 shows the stability values for which the ESM yields better fits than the best performance achieved by any other model. The performance of the ESM is moderately sensitive to stability and decreases smoothly both above and below the optimum of 0.53. We also estimated prediction error for each model using cross-validation, which reduces the risk of over-fitting and provides a more accurate assessment of a model's performance. The results of this analysis are reported in Appendix C, and show similar sum squared error for the ESM (0.29 versus 0.28).

The SM performs substantially worse than the other models overall, and one possible reason is that the SM model does not allow for random guessing and other kinds of noisy responses. The MNM has a parameter than explicitly provides for guessing, and the ESM is tolerant of noise because it naturally favors less-extreme probabilities than the SM. To compare the models on more equal terms, we evaluated a version of the SM that fitted a guessing parameter analogous to the one used in the MNM. The results are reported in Appendix C, which shows that the SM with guessing has much smaller errors than the original version, but that these errors are still more than twice the corresponding errors for the ESM and the MNM.

Individual-level model fits

It is possible for a model to account for mean judgments while failing to capture the behavior of individuals. To explore whether the ESM accounts for the inferences of individuals, we analyzed the correspondence between each model's predictions and individual judgments, using the same model parameters as before. Figure 14 shows the number of individuals whose judgments across all experiments are best matched by each model. More than half of all participants' judgments were best predicted by the ESM, with the remaining participants distributed among the alternative models. We can therefore conclude that the predictive accuracy of the ESM is not an artifact of averaging over individuals.

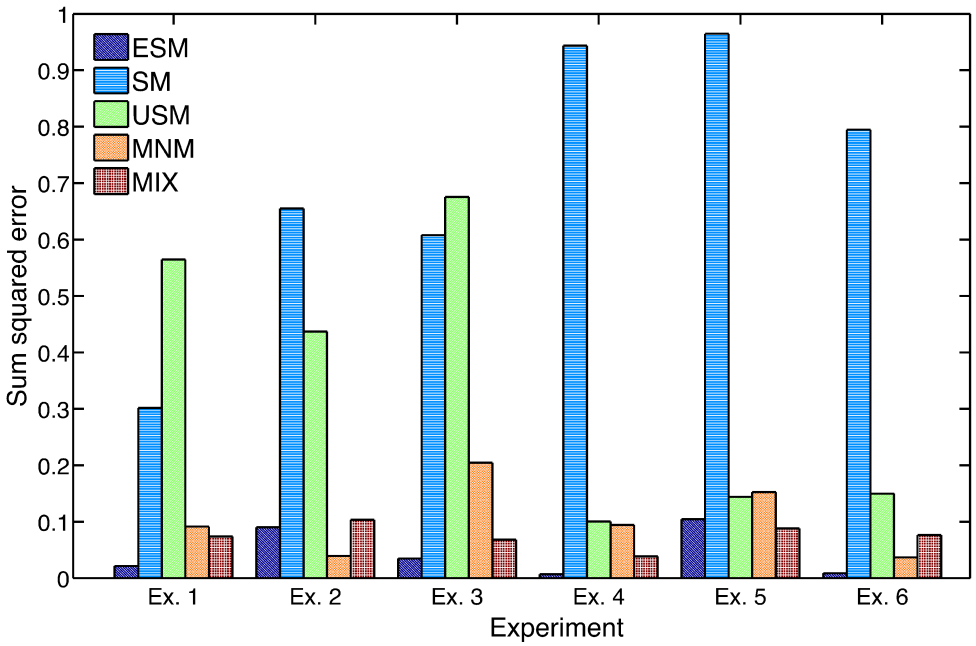


Figure 11. Errors of different models on a per-experiment basis, as measured by sum squared error of predictions versus average judgments. The mixture model is a weighted combination of the USM and the SM.

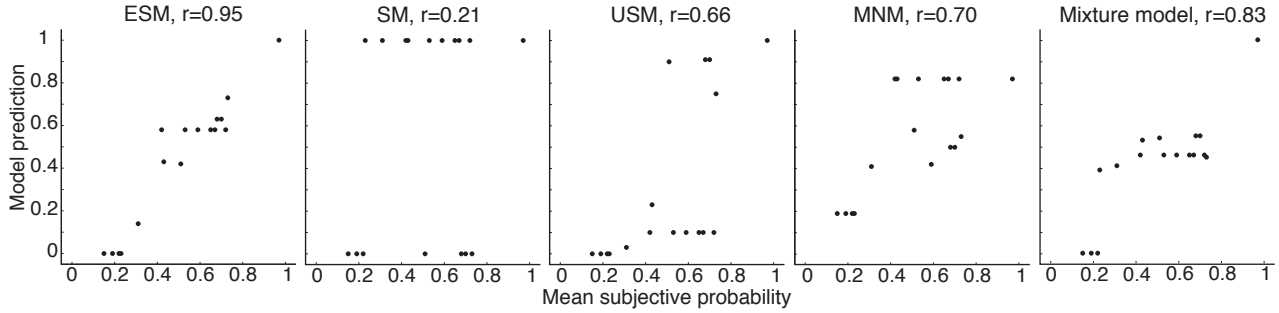


Figure 12. Model predictions plotted against mean human judgments, across all experiments and questions, excluding manipulation-check questions about the counterfactual premises themselves.

Model fits to previous data

We have seen that the accuracy of the ESM is not an artifact of averaging over individuals, but how sensitive is it to the stimuli that we used? We can address this question by looking at how well the ESM predicts human judgments in past studies. Rips (2010) conducted four experiments that are well-suited to this purpose. These experiments were intended to evaluate the Minimal Networks Model and used a cover story in which the variables were the states (operating or not) of different components in a machine. In Rips's Experiment 1, components A and B caused component C to operate either "usually" or "always", and were either jointly necessary or individually sufficient to produce C's operation, for four conditions in all. Experiment 2 looked at effects of question wording—in particular, at the difference between phrases like "if Component C were not operating?" and "if Component C had not operated?" Neither the ESM nor the MNM distinguishes between these alternative wordings. Experiment 3 compared high (.95) and low (.05) base rates for the operation of components A and B, which were left unstated in Experiments 1 and 2. Experiment 4 explored both forward and backtracking counterfactual inferences when A caused B and C, B caused C, and the counterfactual premise focused on B.

In generating predictions for these studies, we took "usually" to indicate a probability of 0.8, and "always" to indicate a probability of 1.0. When base rates were unstated, we assumed that they were 0.5. We compared the ESM to the MNM by finding the parameter values for both models that maximized correlations between the models and mean hu-

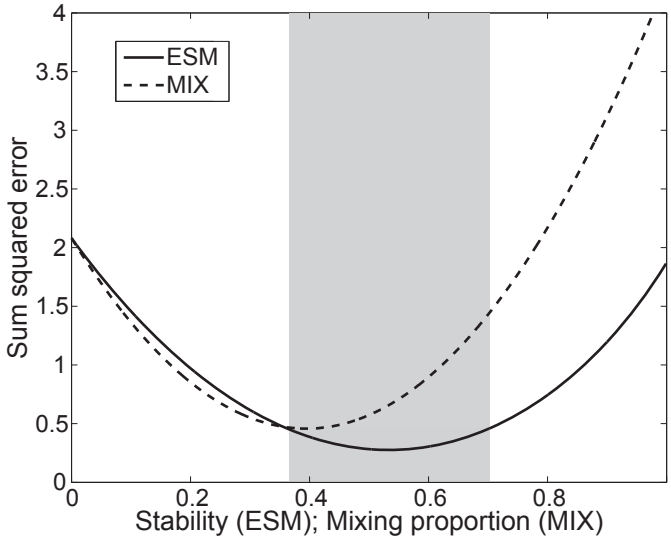


Figure 13. Sensitivity of model fits to parameter choices, comparing ESM and its stability parameter to the mixture model and the weight it assigns to the SM. Parameter values are plotted against sum squared error over all experiments. The shaded region represents the range of values for which the ESM offers more accurate predictions than any mixture of the SM and USM.

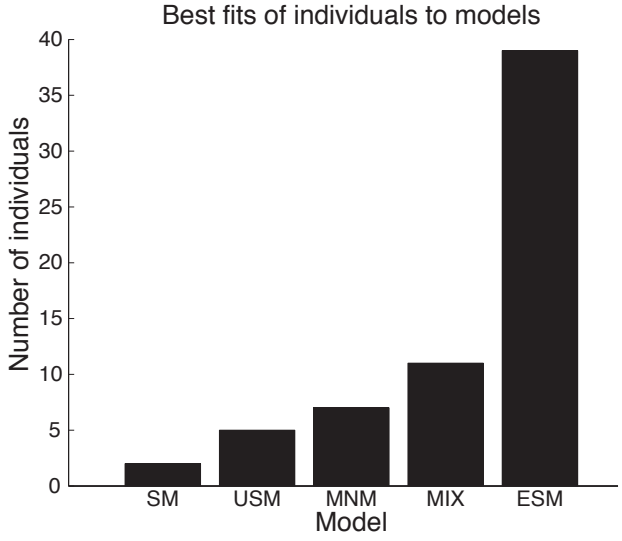


Figure 14. Numbers of individual participants who are best fit by each model, as determined by lowest sum squared error over all experimental conditions.

man judgments from Rips's Experiments 1, 2, 3, and 4. Full details of the analysis are provided in Appendix C. Figure 15 suggests that the ESM accounts for Rips' data as well as his own Minimal Networks Model ($r = .86$ for the ESM, versus $r = .85$ for MNM). The two models perform similarly even though the ESM has one free parameter and the MNM has two.

Taken together, our analyses suggest that the ESM predicts both aggregate and individual-level performance across a variety of counterfactual scenarios. The model performs better than all alternatives that we have considered, including the SM, the MNM, and the mixture model. In addition to the quantitative results that we have presented, our experiments provide qualitative support for all four of the key commitments of the ESM that are highlighted in Table 1. In particular, our data suggest that people are willing to consider counterfactual scenarios beyond those that depart in minimal respects from the actual world, that they tend to assume that apparently stochastic relationships are mediated by hidden variables, that they distinguish between counterfactual observations and interventions, and that they readily backtrack when reasoning about counterfactual scenarios.

General discussion

We presented a formal model of counterfactual reasoning and evaluated it in six experiments. Experiments 1 through 3 provided support for the core commitments of the ESM. Experiments 1 and 2 demonstrated that people are moderately inclined to preserve real-world states of variables when reasoning about counterfactual scenarios. In addition, Experiment 2 showed that people distinguish between counterfactual observations and counterfactual interventions. Experiment 3 demonstrated that people appear to think of stochastic relationships as being mediated by hidden variables, and make inferences consistent with preserving those variables' states.

Experiments 4 through 6 were specifically designed to contrast the ESM with alternative accounts of counterfactual reasoning. Experiments 4 and 5 focused on the Minimal Networks Model, and suggested that people rely on probabilistic information in a way that is consistent with the ESM but inconsistent with the Minimal Networks Model. Experiment 6 focused on a case in which the ESM accounts for human inferences better than any weighted combination of the other models considered in this paper.

Having demonstrated that the ESM predicts human performance in a variety of counterfactual scenarios, we now address some of the general questions that are raised by our approach.

Stability and mutability

The ESM relies on the notion of stability to capture how closely a counterfactual world should be coupled to the real

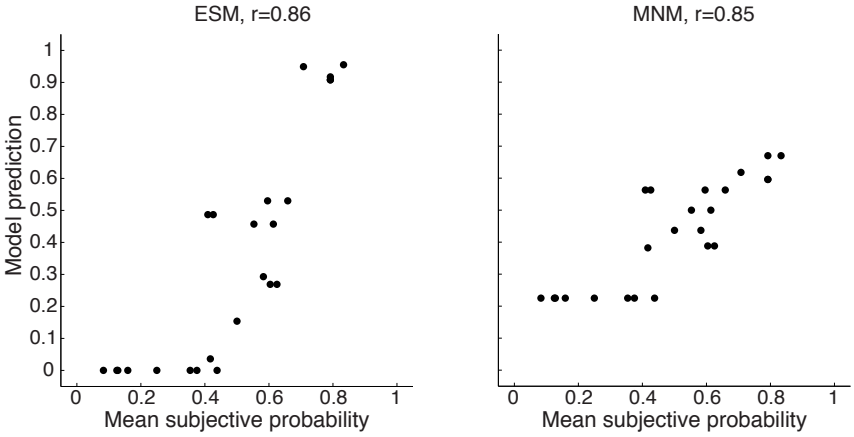


Figure 15. Model predictions for the ESM and MNM, plotted against mean human judgments from Rips (2010).

world. The stability parameter s in Equation 1 can be interpreted in at least two ways. At a process level, s can be viewed as a parameter that influences how a representation of a counterfactual world is constructed. Each exogenous variable in the counterfactual world can be set in two ways: the value of the variable can be copied over from the real world, or the value can be sampled from the prior distribution for this variable. The stability parameter can be thought of as the probability that the value of any given exogenous variable will be copied from the real world rather than sampled from the prior.

Although operationalizing stability at a process level may be sufficient for some purposes, one might also ask why reasoners use the process just described to construct their representation of a counterfactual world. As suggested earlier, the ESM is consistent with the view that a branching process operates through time to create a multiverse of possible worlds. Appendix A describes how the stability parameter can be formalized as a decreasing function of the time that has elapsed since the counterfactual world separated from the real world.

The derivation in Appendix A should be viewed as just one way in which the stability parameter can be understood and in which Equation 1 can be justified. Alternate or expanded derivations will be needed to account for all of the contexts in which counterfactual statements can be evaluated. For example, the derivation in Appendix A applies most naturally to counterfactuals involving premises that refer to particular times. Some counterfactuals, however, do not fit this pattern, including the claim that "if kangaroos had no tails, they would topple over" (Lewis, 1979). The premise of this counterfactual does not refer to a particular time, which means that any derivation involving a temporal branching process seems unlikely to apply to this example.

Our notion of stability is closely related to the notion of mutability that plays a prominent role in previous work on counterfactual reasoning (Kahneman & Miller, 1986; Dehghani et al., 2012). Mutability and stability can be viewed as opposites: for example, an exogenous variable is *mutable* if it is likely to vary across real and counterfactual worlds, and *stable* if it is likely to take identical values across these worlds. Previous studies have identified multiple factors that influence mutability. For example, Kahneman and Miller suggest that exceptional events are more mutable than routine events, that events about which little is known are relatively mutable, that effects are more mutable than causes, and that the second member of an ordered pair of events is more mutable than the first (see also Segura et al., 2002). Girotto, Legrenzi and Rizzo (1991) report that controllable events (e.g. voluntary choices) are more mutable than uncontrollable events (e.g. an asthma attack). Expanding on this result, Mandel and Lehman (1996) suggest that controllable events that prevent a focal outcome are more mutable than controllable events that generate the same outcome.

Mutability is typically described as a property of individual variables, and our notion of stability is also naturally interpreted in this way. As described earlier, our evaluation of the ESM made the simplifying assumption that the stability parameters associated with all exogenous variables were identical. We made this assumption because our experiments focused on simple scenarios in which all variables were of the same type (all represented the presence or absence of hormones) and in which factors such as controllability played little role. It is straightforward, however, to associate different exogenous variables with different values of the stability parameter s in Equation 1. Making this change allows the ESM to incorporate many of the findings that have emerged from previous studies of mutability. For example, exogenous variables associated with uncontrollable events can be assigned lower stability parameters than exogenous variables associated with controllable events.

Setting stability parameters on the basis of prior research provides one way to connect our work with previous studies

of mutability, but does not itself explain the findings of these studies. It seems likely, however, that our formal framework can help to account for at least two findings from the literature on mutability. First, the ESM balances the prior probability of a counterfactual world against consistency with the actual world, and incorporating prior probability may help to explain why exceptional events are more mutable than routine events. Second, Appendix A formalizes stability in a way that relies on a temporal branching process, and it seems likely that this approach can be adjusted to capture the idea that recently-occurring events are more mutable than events that occurred in the distant past. We suspect, however, that many other factors that affect mutability are unlikely to emerge from our framework. For example, explaining why controllable events are more mutable than uncontrollable events seems to require ideas that go beyond the scope of the ESM.

Interpreting counterfactual premises

One of the features of the ESM highlighted in Table 1 is that it distinguishes between counterfactual observations and counterfactual interventions. Counterfactual claims, however, are often expressed using generic statements such as "If A had been absent" which do not indicate whether the absence of A should be treated as an observation or as the result of an intervention. All of our analyses assumed that generic counterfactual premises should be interpreted as observations, and this assumption is supported by our finding that generic counterfactuals typically led to backtracking. In general, however, the interpretation of a counterfactual premise may be influenced by various contextual factors.

Rips and Edwards (2013) demonstrated that relatively subtle variations in the wording of a counterfactual premise can affect the way in which it is interpreted. In their first experiment they asked participants to reason about devices with multiple components, and used two kinds of counterfactual premises. Some participants were asked questions such as "If component B had not operated, would component A have operated?", and others were asked questions such as "If component B had failed, would component B have operated?" Their data suggested that the *not operated* wording led participants to treat the premise as a counterfactual observation, and that the *failed* wording led participants to treat the premise as a counterfactual intervention. Explaining effects of this kind seems to require a careful account of the semantics of the verbs in question.

Pragmatic factors may also shape the interpretation of counterfactual premises. People tend to expect that statements will be informative and relevant, and may therefore reject interpretations under which the counterfactual premise carries little or no information about the answer to the query. For example, suppose that A's presence causes B to be present, which in turn causes C to be present, and that in reality all three variables are present. A participant is asked "If B had been absent, would A have been present or absent?" Interpreting the premise as a counterfactual intervention may violate pragmatic expectations because a premise of this sort provides no information about the value of A. The interventional interpretation, however, appears more natural if the question about A follows a question about C. In this case, the inference that interventions on B provide no information about A seems more acceptable because it contrasts with the inference that interventions on B do provide information about C. This pragmatic account is supported by experiments which demonstrate that people's tendency to backtrack depends on whether backtracking questions are asked before or after non-backtracking questions (Gerstenberg, Bechlivanidis, & Lagnado, 2013).

Applications of counterfactual reasoning

A successful account of counterfactual reasoning should provide a foundation for understanding the role that counterfactual thinking plays in everyday life. This section describes two ways in which the ESM may serve as a useful starting point for studying applications of counterfactual reasoning.

Counterfactuals are useful in part because they provide a guide to action. For example, a student who fails an exam might be interested in actions that she can take to avoid failing future exams. One way to identify these actions is to imagine a counterfactual world in which the student passed the exam, and to reason about the causes that might have generated this positive outcome. For example, the student might think "in order to pass, I'd have needed to study twice as hard as I did." An inference of this kind seems closely related to a backtracking counterfactual that proceeds from a desired effect (e.g. passing the exam) to a cause (e.g. studying hard). Philosophers sometimes suggest that backtracking counterfactuals are invalid or treat them as anomalous special cases (Downing, 1959; Lewis, 1973a; Bennett, 1974), but these counterfactuals may be relatively common when planning future actions. Unlike the SM, the ESM supports backtracking, and therefore seems well-suited for capturing the role that counterfactuals play in action selection.

Assigning credit and blame is a second task that often draws on counterfactual reasoning. For example, a student who fails an exam is likely to blame herself for this outcome if she thinks that she would have needed to study more in order to pass. A different student might think that "I would have passed if the test had been easier," and might be inclined to blame her failure on the difficulty of the exam. As these examples suggest, credit and blame assignment depend critically on the mutability of different variables (Miller & Gunasegaram, 1990; Alicke, Buckingham, Zell, & Davis, 2008)—for example, whether the preparation of the student or the difficulty of the exam is the more mutable factor. As suggested earlier, the prior distributions used by the ESM can

capture some of the ways in which norms influence mutability (Kahneman & Miller, 1986). For example, suppose that the student normally prepares well for exams, and that these exams are normally difficult. Given prior distributions that capture these expectations, the ESM predicts that the student who failed is likely to blame herself for this outcome, because the counterfactual premise "I passed the exam" provides more support for the conclusion "I prepared well" than for the conclusion "the exam was easy." As mentioned earlier, factors including temporal sequence and the controllability of different events can also influence mutability judgments. Allowing different exogenous variables to take different values of the stability parameter may allow the ESM to capture more of the ways in which mutability influences credit and blame assignment.

Levels of analysis

The psychological literature includes contributions at multiple levels of analysis (Marr, 1982), including process models that specify the sequence of steps used to solve a problem, and "computational-level models" that specify how a problem would be solved by an ideal observer. This section argues that the ESM is relevant to both families of models.

We suggested earlier that the computations specified by the ESM can be carried out using an "extended twin network" like the example shown in Figure 2c. An extended twin network is a special case of a Bayes net, which means that arguments about the underlying algorithms and neural implementations for other Bayes net accounts (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Gopnik et al., 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) apply to the ESM as well. To convert a Bayes net into a fully-specified process model, the network must be supplemented with an algorithm for carrying out inference over the network. The standard algorithms for inference in Bayes nets make use of local message-passing along the edges in the network, and can therefore be viewed as instantiations of the psychological notion of spreading activation (Pearl, 1994).

Inference over extended twin networks provides one way to compute the predictions of the ESM, but these predictions could also be computed by sampling possible worlds from the distribution specified by the ESM. Sampling has been widely used to develop process-level accounts of probabilistic inference, and the same general-purpose sampling methods that have previously been applied to problems including categorization, decision-making and sentence processing can also be applied in our setting (Griffiths, Vul, & Sanborn, 2012). A sampling account of counterfactual reasoning is appealing in part because it captures the intuition that people consider only a small handful of concrete possibilities when asked to evaluate a counterfactual claim (Byrne, 2002).

As just suggested, the ESM appears closely related to previous process models of human reasoning, but our approach can also be viewed as a contribution at Marr's "computational level". In particular, Appendix A suggests that the ESM captures the "right" way to make counterfactual inferences assuming that possible worlds are generated from the branching process described in the appendix. Metaphysical assumptions of this kind seem to lie beyond the reach of normative justification, but given the generative assumptions in Appendix A, the rest of our theory follows from the dictates of rational probabilistic inference.

Alternative accounts of counterfactual reasoning

Throughout we have compared multiple models of counterfactual reasoning, but all of these models rely on probabilistic inference over causal networks. This section discusses how our work relates to alternative accounts of counterfactual reasoning.

Possible-worlds approaches. As mentioned earlier, an influential line of work proposes that counterfactual arguments are evaluated by considering possible worlds that make the premises true and that are as close as possible to the real world (Lewis, 1973a; Stalnaker, 1987). This possible-worlds approach has subsequently influenced several psychological theories of conditional reasoning (Rips & Marcus, 1977; Evans & Over, 2004; Oaksford & Chater, 2010a). Possible-worlds accounts can be formalized using *imaging*, which is a procedure for adjusting probability distributions in response to a premise (Lewis, 1976). This section explains how the ESM can be characterized in terms of an imaging operation over possible worlds.

In our setting, each possible world corresponds to a setting of all of the variables (both exogenous and endogenous) in a functional causal model. For example, in our bacon-cooking scenario each possible world can be characterized by a vector that assigns values to all six variables in the functional model in Figure 1b. A reasoner may have knowledge K that specifies the values of some variables, but in general she will be uncertain about others, and her beliefs can be captured using a distribution $P(W|K)$ that captures her uncertainty about the true state of the world. For example, suppose that the reasoner observes that bacon is not cooked, that the alarm does not activate, and that the neighbors are not disturbed. In this case $P(W|K)$ assigns nonzero probability to two possible worlds: one in which cooking bacon would have activated the alarm (i.e. $U_S = t$) and another in which bacon would not have activated the alarm ($U_S = f$). In both of these worlds, the remaining variable assignments are $B = f$, $S = f$, $N = f$, $U_B = f$, and $U_N = f$.

Suppose that the reasoner now considers the counterfactual conditional "if bacon had been cooked then the neighbors would have been disturbed." Imaging is a procedure for adjusting the distribution $P(W|K)$ in light of the premise $B = t$ to create a new distribution $P_B(W|K)$ over worlds (Lewis, 1976). The adjusted distribution assigns nonzero probability

only to worlds that make the premise $B = t$ true, and adopting this distribution is therefore a way to determine what would follow from accepting the premise. In the most basic case, the adjusted distribution $P_B(W|K)$ is generated by shifting the probability $P(W = w|K)$ assigned to world w to the closest world w_B that makes the premise true. For example, consider the world in which conditions were right for B to trigger the alarm, but B did not occur ($B = f$, $S = f$, $N = f$, $U_B = f$, $U_S = t$, $U_N = f$). The probability associated with this world might be shifted to a world in which B does indeed activate the alarm ($B = t$, $S = t$, $N = t$, $U_B = t$, $U_S = t$ and $U_N = f$). The probability of the counterfactual conditional "if B then N" is then defined as $P_B(N = t|K)$, or the probability that N is true according to the adjusted distribution $P_B(W|K)$.[5]

The imaging procedure just described makes use of a similarity measure over possible worlds, but does not specify exactly how this measure should be defined. Pearl's SM is a special case of imaging that uses functional causal models to characterize the world w_B that is closest to world w (Pearl, 2000). According to the SM, all exogenous variables in w_B take the same values that they had in w, and all remaining variables are deterministically specified using a functional causal model that reflects an intervention that makes the premise $B = t$ true.

As characterized so far, imaging assumes that every world w shifts all of its probability to a unique world w_B that is closer to w than are all other worlds that make the premise true. Gärdenfors (1988) relaxes this assumption by allowing a world w to divide its probability among multiple worlds that make the premise true. He motivates this generalization by suggesting that the probability of w can be divided among several worlds that are "equally close" to w. His formalization of *general imaging*, however, allows the probability of w to be divided among all worlds that make B true, including those that are both close to and far from w (Gärdenfors 1988, p. 112).

Our new model can be viewed as a special case of general imaging. Consider again our bacon example, and suppose that we are using an extended twin network to reason about the counterfactual conditional "if B then N," where the premise B is treated as a counterfactual observation. The extended network will be identical to Figure 3a, except that Figure 3a shows a counterfactual premise involving S rather than B. The marginal distribution $P(W|K)$ over the left half of the extended network captures uncertainty about the true state of the world, and the marginal distribution over the right half of the network can be interpreted as an adjusted distribution $P_B(W|K)$ produced by a general imaging operation. This operation has the property that each world w divides its probability among *all* worlds that are consistent with the counterfactual premise. The division tends to favor worlds that are both similar to w and probable *a priori*, and the stability parameter s specifies the tradeoff between similarity and prior probability. In this context, similarity should be interpreted as the extent to which two worlds assign the same values to the exogenous variables.

Previous authors have pointed out that imaging can be sensitive to both similarity and prior probability (Joyce, 2010; Pearl, 2010). Joyce (2010, p. 149) suggests that imaging should involve "the *combined* effects of judgments of similarity among worlds and prior probabilities," and describes a procedure called *Bayesian imaging* that has this property. Under Bayesian imaging, each world w transfers its probability to a set of worlds that are all equally close to w, and the amount of probability received by each world in this set is proportional to its prior probability. Bayesian imaging can therefore be viewed as a two step process: first apply similarity to filter out all worlds except those that are close to w, and then use prior probability to decide how the probability of w should be distributed over the worlds that remain. This two-step characterization highlights the fact that Bayesian imaging privileges similarity over prior probability: w can transfer its probability to worlds with both high and low prior probabilities, but can only transfer its probability to worlds that are maximally similar to w. The ESM treats similarity and prior probability in a more balanced way, and allows w to shift its probability to worlds that are both high and low in prior probability, and both close to and far from w.

Our comparison between Bayesian imaging and the ESM highlights the fundamental way in which our approach differs from previous possible-worlds approaches. To our knowledge, all of these approaches make the minimality assumption described earlier: they assume that counterfactual arguments should be evaluated by considering worlds that depart in minimal respects from the actual world. As we have discussed, Pearl's SM and the Minimal Networks Model make the same assumption. In contrast, the ESM works with a distribution over counterfactual worlds that assigns non-zero probability to worlds that are both close to and far from the actual world. Although non-standard, our approach is supported by the data that we have presented. In particular, Experiments 4 and 5 were specifically designed to ask whether people consider only possible worlds that differ in minimal respects from the actual world, and the results raise challenges for any formal account that relies on the minimality assumption.

Connectionist approaches. Although the probabilistic approach provides a natural way to incorporate uncertainty and make graded predictions, the connectionist approach of-

[5] $P_B(N|K)$ is different in general from the conditional probability $P(N|B)$. For the example in the text, imaging produces $P_B(N = t|K) = 0.9$, but $P(N = t|B = t) = 0.91$. The two quantities differ because the latter allows for possible worlds in which $U_N = t$ and the neighbors are disturbed even though the alarm does not activate. These worlds do not contribute to the imaging computation because knowledge K rules out the possibility that $U_N = t$.

fers the same advantages. Oaksford and Chater (2010b) have developed a connectionist model that captures some aspects of their probabilistic theory of conditional reasoning, and it may also be possible to develop a connectionist implementation of the ESM. At present, however, connectionist models have not been shown to capture the kinds of counterfactual inferences considered in this paper.

To our knowledge, the only existing connectionist model of counterfactual reasoning is the μKLONE model developed by Derthick (1987). Derthick's model treats counterfactual conditionals as situations in which one fact — the counterfactual premise — must be reconciled with other incompatible facts, namely the state of the real world. To achieve this reconciliation, the model relies on a large number of hand-coded relationships and constraints, and combines logical and connectionist representations to find configurations of variables that minimize the strength and number of constraints being violated. Importantly, however, the model is not natively equipped with the capacity to reason about causal systems like those in our experiments. It may be possible to apply μKLONE to our experiments by supplying it with a set of rules and constraints that capture the causal relations and base rates in our scenarios, along with abstract principles such as the difference between observation and intervention. Developing these rules and constraints, however, is a research challenge that goes substantially beyond the analyses presented by Derthick.

Conclusions

We presented a probabilistic model, the ESM, which extends and generalizes Pearl's account of counterfactual reasoning. Like Pearl's account, the ESM does not treat causal relationships as being inherently stochastic, but instead explains unpredictable phenomena in terms of exogenous variables. Unlike Pearl's account, the ESM allows the values of these exogenous variables to differ across real and counterfactual worlds. Treating exogenous variables in this way allows the ESM to go beyond Pearl's approach by accounting for backtracking counterfactuals and capturing the difference between counterfactual interventions and counterfactual observations.

Although we have argued that the ESM improves on previous models of counterfactual reasoning, all of the models evaluated in this paper are similar in one important respect: all make use of Bayesian networks. The Bayes net approach is a general framework that has been applied to many aspects of causal cognition, including causal learning, causal prediction and causal explanation. In all of these cases, causal cognition is characterized in terms of probabilistic inference over graph-structured representations. Our work endorses the view that the same cognitive machinery supports inferences about counterfactual states of affairs.

References

Adams, E. W. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, *6*(1), 89–94.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403–418.

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin*, *34*(10), 1371–1381.

Bennett, J. F. (1974). Counterfactuals and possible worlds. *Canadian Journal of Philosophy*, *4*(2), 381–402.

Bennett, J. F. (2003). *A philosophical guide to conditionals*. Oxford: Clarendon Press.

Byrne, R. M. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, *6*(10), 426–431.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55-85.

Derthick, M. (1987). Counterfactual reasoning with direct models. In *Proceedings of the Sixth National Conference on Artificial Intelligence* (Vol. 1, pp. 346–351).

Downing, P. B. (1959). Subjunctive conditionals, time order, and causation. In *Proceedings of the Aristotelian society* (pp. 125–140).

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Epstude, K., & Roese, N. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*(2), 168–192.

Evans, J., & Over, D. (2004). *If*. Oxford: Oxford University Press.

Fernbach, P., & Erb, C. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1327–1343.

Frosch, C. A., & Johnson-Laird, P. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*(3), 280–291.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2386–2391).

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1), 111–133.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565-610.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in

children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In L. Schulz & A. Gopnik (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Halpern, J. Y., & Pearl, J. (2001). Causes and explanations: A structural-model approach – Part I: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (p. 194-202). San Francisco, CA: Morgan Kaufmann.

Harper, W. L. (1981). A sketch of some recent developments in the theory of conditionals. In W. L. Harper, R. Stalnaker, & G. A. Pearce (Eds.), *Ifs* (pp. 3–38). London: D. Reidel Publishing Company.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755.

Hume, D. (1748). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett.

Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, *55*(1), 3–21.

Joyce, J. M. (2010). Causal reasoning and backtracking. *Philosophical Studies*, *147*(1), 139–154.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136.

Karlin, S., & Taylor, H. (1975). *A first course in stochastic processes* (2nd ed.). San Diego, CA: Academic Press.

Kemp, C., Shafto, P., & Tenenbaum, J. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1), 35–73.

Lewis, D. (1973a). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (1973b). *Counterfactuals*. Malden, MA: Basil Blackwell.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, *85*, 297–315.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Luhmann, C. C., & Ahn, W. (2005). The meaning and computation of causal power: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review*, *112*(3), 685–693.

Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, *71*(3), 450.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

McCall, S. (1984). Counterfactuals based on real possible worlds. *Nôus*, *18*(3), 463–477.

McDermott, M. (2007). True antecedents. *Acta Analytica*, *22*(4), 333–335.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249–264.

Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology*, *59*(6), 1111–1118.

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, USA.

Oaksford, M., & Chater, N. (2010a). Causation and conditionals in the cognitive science of human reasoning. *Open Psychology Journal*, *3*, 105–118.

Oaksford, M., & Chater, N. (2010b). Conditional inference and constraint satisfaction: Reconciling mental models and the probabilistic approach. *Cognition and conditionals: Probability and logic in human thinking*, 309–333.

Over, D. E., Hadjichristidis, C., Evans, J. S. B., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.

Pearl, J. (1994). Belief networks revisited. *Artificial intelligence in perspective*, 49–56.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Pearl, J. (2010). *Physical and metaphysical counterfactuals* (Tech. Rep. No. R-359). CA: Department of Computer Science, University of California, Los Angeles.

Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive Science*, *37*(6), 977–985.

Pearl, J., & Bareinboim, E. (2011). *Transportability across studies: A formal approach* (Tech. Rep.). Los Angeles, CA: DTIC Document.

Ramsey, F. P. (1978). General propositions and causality.

In D. H. Mellor (Ed.), *Foundations: essays in philosophy, logic, mathematics and economics* (pp. 133–151). London: Routledge and Kegan Paul.

Reichenbach, H. (1956). *The direction of time*. Berkeley, CA: University of California Press.

Richardson, T. S., & Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences working paper series*.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–1135.

Rips, L. J., & Marcus, S. (1977). Suppositions and the analysis of conditional sentences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 185–220). Hillsdale, NJ: Lawrence Erlbaum Associates Hillsdale.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. M. (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, *77*(2), 427-442.

Segura, S., Fernandez-Berrocal, P., & Byrne, R. (2002). Temporal and causal order effects in thinking about what might have been. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(4), 1295–1305.

Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. *The Journal of Machine Learning Research*, *9*, 1941–1979.

Shpitser, I., & Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 514–521).

Singmann, H., Klauer, K. C., & Over, D. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, *5*(316).

Sloman, S., & Lagnado, D. (2005). Do we 'do'? *Cognitive Science*, *29*(1), 5–39.

Slote, M. A. (1978). Time in counterfactuals. *The Philosophical Review*, *87*(1), 3–27.

Spellman, B., & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems*, *64*(4), 241–264.

Spellman, B., Kincannon, A., & Stose, S. (2005). The relation between counterfactual and causal reasoning. In D. R. Mandel, D. J. Hilton, & P. Catellani (Eds.), *The psychology of counterfactual thinking* (pp. 28–43). New York, NY: Routledge.

Stalnaker, R. C. (1981). A theory of conditionals. In W. L. Harper, R. Stalnaker, & G. A. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance and time* (pp. 41–55). Springer.

Stalnaker, R. C. (1987). *Inquiry*. Cambridge, MA: MIT Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental and computational approaches* (pp. 167–204). Cambridge: Cambridge University Press.

Unterhuber, M. (2013). *Possible worlds semantics for indicative and counterfactual conditionals?: A formal philosophical inquiry into Chellas-Segerberg semantics*. Frankfurt: Ontos Verlag.

Walters, L. (2009). Morgenbesser's coin and counterfactuals with true components. In *Proceedings of the Aristotelian society* (Vol. 109, pp. 365–379).

Woodward, J. (2011). Psychological studies of causal and counterfactual reasoning. *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*, 16.

Zeelenberg, M., van Dijk, W. W., Manstead, A. S., & van der Pligt, J. (2000). On bad decisions and disconfirmed expectancies: The psychology of regret and disappointment. *Cognition & Emotion*, *14*(4), 521–541.

Appendix A
Temporal branching and stability

This appendix presents one way in which the ESM's notion of stability can be formalized. We will derive Equation 1 from the assumption that worlds are generated by a branching process that operates through time.

Figure A1 illustrates how this branching process might produce a tree of possible worlds. At any given time, the exogenous variables within any given world are set to specific values, but the values of these variables may evolve over time. For simplicity, we will assume for now that all worlds contain a single exogenous variable called U. Let U^1 and U^2 denote the value of this variable in two worlds, world 1 and world 2. Suppose that the most recent common ancestor of these worlds is world 0, and let U^0 denote the value of variable U in world 0.

When branching events occur, the parent world (e.g. world 0) "splits" into two identical copies (e.g. worlds 1 and 2). Although the two copies are identical immediately after

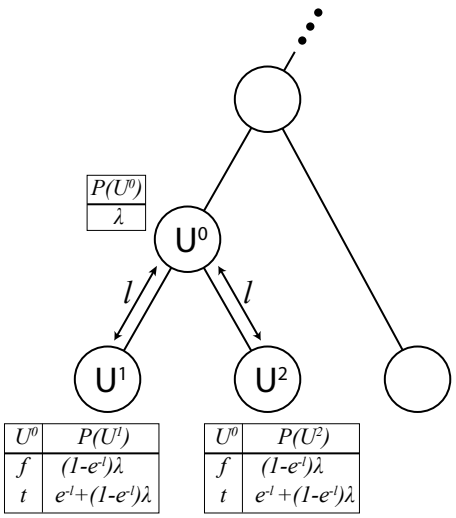


Figure A1. Graphical depiction of a branching process over possible worlds.

the split, after that point the two worlds evolve independently over time. Figure A1 indicates that an interval of length l has elapsed since world 0 split to form worlds 1 and 2. During this interval, variable U has evolved independently within the two worlds, and as a result U^1 and U^2 may take different values.

We can formalize the way in which variables change through time by assuming that the value of a variable at a given instant depends only on the long-run base rate of that variable and the value of that variable at the previous instant of time. In Figure A1, let λ denote the long-run base rate of variable U. Technically speaking, we assume that variable U evolves according to a continuous-time Markov chain with infinitesimal matrix:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ 1-\lambda & -(1-\lambda) \end{bmatrix} \tag{2}$$

The same generative process is widely used as a simple model of biological evolution (Huelsenbeck & Ronquist, 2001), and has also been used to develop psychological accounts of generalization (Tenenbaum, Kemp, & Shafto, 2007; Kemp, Shafto, & Tenenbaum, 2012).

Given these assumptions, we can convert the tree in Figure A1 into a Bayesian network that allows us to compute the value of variable U at different points in the tree. The CPD for U^0 shows that U is true in world 0 with probability λ, which is consistent with our assumption that λ is the long-run base rate of U. Given the value of U^0, the CPDs for U^1 and U^2 show that the values of these variables depend on the length of the interval l that has elapsed since worlds 1 and 2 separated. If l is very small, then U^1 and U^2 will be equal to U^0 with high probability. If l is very large, then the values of U^1 and U^2 are effectively independent of U^0 and are determined only by the base-rate. The CPDs shown in Figure A1 can be derived from our assumption that variable U evolves according to the continuous-time Markov chain specified by Equation 2 (Karlin & Taylor, 1975).

The key component of the ESM is Equation 1 in the main text, and this equation can be derived from the branching process just characterized. If world 1 is the actual world and world 2 is the counterfactual world under consideration, then the distribution in Equation 1 corresponds to $P(U^2|U^1)$. The CPDs in Figure A1 imply that this distribution is

$$P(U^2 = t|U^1) = \begin{cases} (1-e^{-2l})\lambda, & \text{if } U^1 = f \\ e^{-2l} + (1-e^{-2l})\lambda, & \text{if } U^1 = t. \end{cases} \tag{3}$$

If we set $s = e^{-2l}$, then Equation 3 can be written as

$$P(U^2 = t|U^1) = \begin{cases} (1-s)\lambda, & \text{if } U^1 = f \\ s + (1-s)\lambda, & \text{if } U^1 = t, \end{cases} \tag{4}$$

which can be rewritten in a format that matches Equation 1 in the main text:

$$P(U^2|U^1) = s\delta(U^1) + (1-s)P(U^2), \tag{5}$$

where $P(U^2 = t) = \lambda$ and $P(U^2 = f) = 1-\lambda$.

The derivation above demonstrates that the stability parameter s in Equation 1 can be interpreted as a decreasing function of the length of time l that has elapsed since the counterfactual world and the actual world separated. Given this interpretation, Equation 1 is a direct consequence of the generative assumptions specified in this appendix.

Until now we have assumed that the worlds under consideration each contain a single exogenous variable. If there are multiple exogenous variables, we allow these variables to have different baserates λ_i, and assume that these variables evolve independently according to the generative process summarized by Figure A1.

The CPDs shown in Figure A1 can be used to complete the specification of the extended twin network shown in Figure 2c. Figure 2c includes exogenous variables that belong to the real and counterfactual worlds, and the CPD for each of these exogenous variables takes the form shown in Figure A1, where the parameter λ is replaced by the base rate of the variable in question. Given these CPDs, the extended twin network in Figure 2c is a fully specified Bayesian network, and standard algorithms for inference in Bayesian networks can be used to compute the predictions of the ESM.

The extended twin network in Figure 2c can be viewed as a Bayesian network that incorporates both temporal relationships between possible worlds and the causal relationships that exist within each world. Our specification of this network is formally identical to a previous model of generalization that incorporates both similarity relationships between objects and causal relationships between the features

of these objects (Kemp et al., 2012). Critically, however, the work of Kemp et al. focuses on variables that belong to the actual world, whereas the ESM focuses on inferences about a possible world that differs from the actual world.

The derivation in this section is related to previous philosophical accounts of counterfactual conditionals that invoke the notion of a temporal branching process. McCall (1984) proposes that possible worlds are organized into a "continuously-branching tree structure," and that the similarity between two worlds is determined by the "amount of common past they share." Bennett (2003) offers a related proposal, and suggests that the similarity between two worlds can be understood by considering the "fork" at which the two diverged. This appendix can therefore be viewed as an attempt to formalize some intuitions about similarity that have previously appeared in the philosophical literature.

Appendix B
Functional causal models

As described in the main text, the ESM and the SM both rely on functional causal models (FCMs). To compute the predictions of these approaches, we used FCMs that matched the descriptions of the causal systems that participants read.

In Experiments 1 and 4, the FCMs can be recovered immediately from the causal graphs in Figure 4a and Figure 7a, respectively. In Experiment 3, the FCM can be recovered from the causal graph in the *deterministic* condition in Figure 6a. For Experiments 2, 5, and 6, the causal graphical models shown in Figures 6, 9 and 10 do not fully capture the FCMs implied by the descriptions that participants saw. The FCMs for these three experiments are shown in Figure B1.

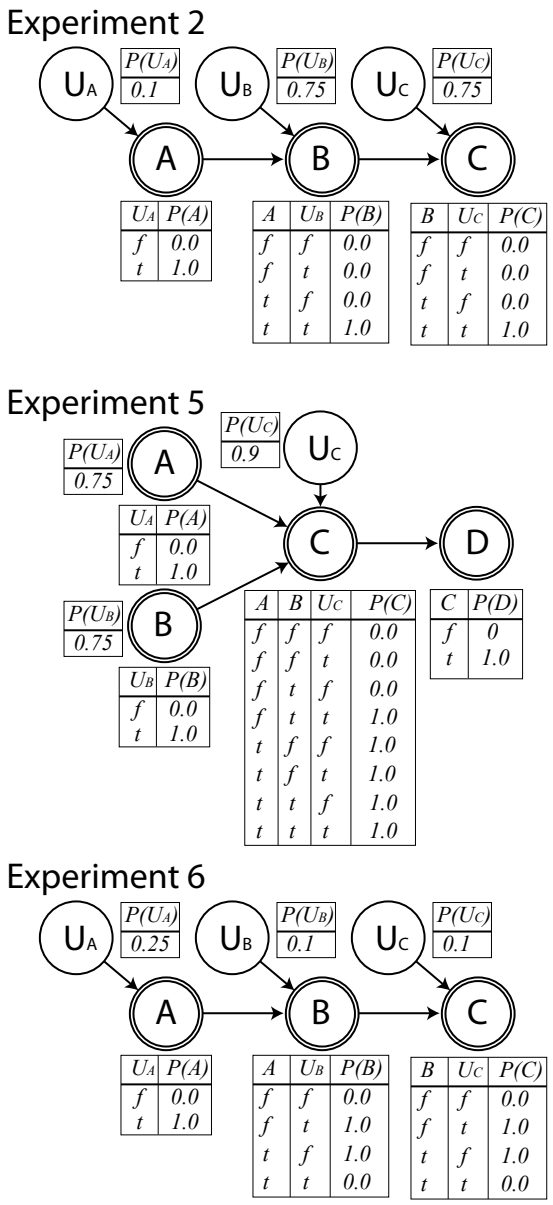


Figure B1. Graphical representations of the functional causal models for Experiments 2, 5, and 6. For Experiment 5, nodes representing the exogenous causes of hormones *A* and *B* have been omitted in the interest of compactness.

Appendix C
Parameters and model fits

Pearl's SM and the USM have no free parameters, and the ESM has a single stability parameter. The MNM has two free parameters: the first captures the rate at which participants randomly guess when making their judgments, and the second captures the weight that they assign to probabilistic evidence. For each model with free parameters, we used one set of parameters to model judgments across all six experiments. The parameters chosen were those that minimized the total sum squared error of the model relative to the mean subjective probabilities provided by participants. In addition to the models described in the main text, we also considered two variants of the SM. The SM-g allows for random guessing by fitting a single "guessing rate" parameter to the data. The SM-go handles default counterfactuals and counterfactual interventions in the same way as the SM-g, but treats counterfactual observations as simple observations (and therefore handles counterfactual observations in the same way as the USM).

In our analysis of Rips's (2010) data, we used parameter values that maximized correlation across mean judgments for Rips's Experiments 1, 2, 3, and 4, minus conditions in Experiment 3 where participants were asked whether outcomes "necessarily follow" from a counterfactual premise. Among the parameter values that maximized correlation, we chose those that produced the lowest sum squared error. For the ESM, the best-fitting stability parameter was 0.68, while for the MNM the best fits were associated with a guessing probability of .45 and a probability of using probabilis-

tic information of .41 in non-guesses. The values of the MNM parameters that minimized error across our own experiments were .37 and .30, for guessing and probabilistic-information parameters, respectively. Maximizing correlations alone would have penalized the MNM by failing to optimize the MNM's guessing parameter, because that parameter's value does not influence correlation values unless it is set to 1.

Because all parameter spaces were at most two-dimensional, and all parameters fell between 0 and 1, we used a simple grid search to find parameter values that minimized sum squared error. In the case of one-parameter models (the ESM, SM-g, SM-go, and mixtures of the USM and SM), we considered 1000 uniformly spaced parameter values between 0 and 1. In the case of two-parameter models (the MNM), we considered 200 values for each parameter, uniformly spaced between 0 and 1, or 40,000 total combinations.

The main text discusses the sensitivity of our error rates to choices of the stability parameter. We expand on that discussion by reporting relative model performances under cross validation. In cross-validation, a data set is partitioned into a training set, which is used to estimate a model's parameters, and a test set, which is used to evaluate the performance of the model. Fitting parameters and evaluating models using the same data can lead to inflated estimates of model performance, and separating training sets from test sets allows a more accurate comparison of the performance of different models. In our case, we used 9 partitions, with data from one scenario set aside for testing in each round. We used the same fitting procedure that we used for our overall parameter estimates, described above, with results reported in the second row of Table C1. This analysis shows that the good performance of the ESM is not due to testing and fitting on the same data, but that the mixture model and the MNM show degraded accuracy under cross-validation. Because the SM and USM do not have any parameters to fit, their error rates are unchanged.

In the main text, our model fits are based on estimates of subjective probabilities that incorporate both forced-choice responses and reported confidences. We also combined the results of online and in-lab participant groups. Table C1 shows that the performance of the ESM remains high if we look at the different populations (in-lab versus online) separately and if we use only forced-choice judgments. Figure C1 shows scatterplots comparing ESM predictions to participant judgments across all questions if we separate in-lab and online populations.

Appendix D
Baseline results

Experiments 1, 2, and 6 included baseline conditions, which did not frame the questions as counterfactuals and omitted information about the real-world states of the variables. Aside from these changes, the baseline conditions were identical to their counterfactual counterparts. Participants' judgments in these conditions are summarized in Figure D1. See the description of Experiment 1 for additional details about these conditions.

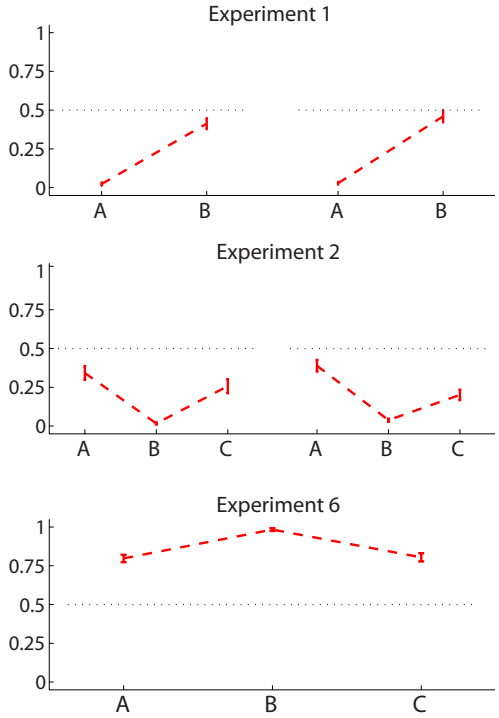


Figure D1. Mean participant judgments for the baseline conditions in Experiments 1, 2, and 6. Error bars represent standard errors of the means.

Appendix E
Experimental materials

This appendix includes the descriptions of the causal systems that were provided to participants and the questions they were asked. For each system, the questions were asked in random order.

Experiment 1, observation condition

> In a very small number of the mice, hormone A is present.
>
> In a very small number of the mice, hormone B is present.
>
> The presence or absence of hormone A does not depend on the status of hormone B.
>
> The presence or absence of hormone B does not depend on the status of hormone A.

Table C1

Model fits as measured by sum squared error across all counterfactual judgments for the Extended Structural Model (ESM), Structural Model (SM), SM with a guessing parameter (SM-g), SM-g that treats counterfactual observations as simple observations (SM-go), Unattached Structural Model (USM), Minimal Networks Model (MNM), and the mixed strategies model (MIX). The first row shows errors for the analysis and populations described in the main text. The second row shows errors using cross-validation, and the remaining rows show errors under different sub-populations and dependent measures; see the text in this Appendix for details.

Group	ESM	SM	SM-g	SM-go	USM	MNM	MIX
Original analysis	0.28	4.28	1.85	1.49	2.08	0.83	0.46
Cross-validation	0.29	4.28	1.91	1.56	2.08	0.92	0.53
Lab participants	0.40	3.65	1.59	1.33	2.70	0.66	0.59
Online participants	0.27	4.93	2.18	1.73	1.64	1.08	0.45
Forced-choice judgments	0.27	4.95	2.56	2.01	1.89	1.18	0.61

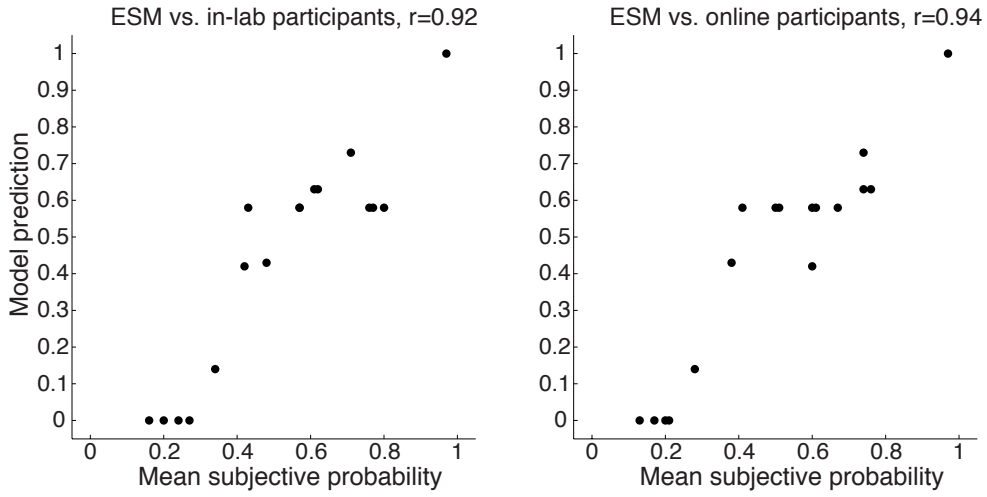


Figure C1. ESM predictions plotted against mean human judgments from in-lab and online participants, across all Experiments and questions, excluding manipulation-check questions about the counterfactual premises themselves.

Fred is a mouse from the Tondam Lab. You test Fred and find that:

Hormone A is definitely present.

Hormone B is definitely present.

For Fred the mouse, if you had observed A to be absent, would A be present?

For Fred the mouse, if you had observed A to be absent, would B be present?

Experiment 1, intervention condition

In a very small number of the mice, hormone A is present.

In a very small number of the mice, hormone B is present.

The presence or absence of hormone A does not depend on the status of hormone B.

The presence or absence of hormone B does not depend on the status of hormone A.

Some mice are raised on regular kibble and others are raised on special kibble.

The two kinds of kibble are identical except for the fact that special kibble contains a substance that specifically prevents the formation of hormone A.

Martha is a mouse from the Runcible Lab raised on regular kibble. You test Martha and find that:

Hormone A is definitely present.

Hormone B is definitely present.

For Martha the mouse, if you had raised Martha on special kibble, would A be present?

For Martha the mouse, if you had raised Martha on special kibble, would B be present?

Experiment 2, observation condition

In almost all of the mice, hormone A is absent.

The status of hormone B is caused by the status of hormone A and nothing else.

In many of the mice, if A is present, this causes B to be present.

The status of hormone C is caused by the status of hormone B and nothing else.

In many of the mice, if B is present, this causes C to be present.

Frank is a mouse from the Nanber Lab. You test Frank and find that:

Hormone A is definitely present.

Hormone B is definitely present.

Hormone C is definitely present.

For Frank the mouse, if you had observed B to be absent, would A be present?

For Frank the mouse, if you had observed B to be absent, would B be present?

For Frank the mouse, if you had observed B to be absent, would C be present?

Experiment 2, intervention condition

In almost all of the mice, hormone A is absent.

The status of hormone B is caused by the status of hormone A and nothing else.

In many of the mice, if A is present, this causes B to be present.

The status of hormone C is caused by the status of hormone B and nothing else.

In many of the mice, if B is present, this causes C to be present.

Some mice are raised on regular kibble and others are raised on special kibble.

The two kinds of kibble are identical except for the fact that special kibble contains a substance that specifically prevents the formation of hormone B.

Nellie is a mouse from the Phroxis Lab raised on regular kibble. You test Nellie and find that:

Hormone A is definitely present.

Hormone B is definitely present.

Hormone C is definitely present.

For Nellie the mouse, if you had raised Nellie on special kibble, would A be present?

For Nellie the mouse, if you had raised Nellie on special kibble, would B be present?

For Nellie the mouse, if you had raised Nellie on special kibble, would C be present?

Experiment 3, stochastic condition

In some of the mice, hormone A is present.

The status of hormone B is caused by the status of hormone A and nothing else.

In a very small number of the mice, if A is present, this causes B to be absent, and if A is absent, this causes B to be present.

In the remaining mice, if A is present, this causes B to be present, and if A is absent, this causes B to be absent.

Ronald is a mouse from the Gostak Lab. You test Ronald and find that:

Hormone A is definitely present.

Hormone B is definitely not present.

For Ronald the mouse, if A were absent, would A be present?

For Ronald the mouse, if A were absent, would B be present?

Experiment 3, deterministic condition

In some of the mice, hormone A is present.

In a very small number of the mice, hormone B is present.

The status of hormone C is caused by the status of hormones A and B and nothing else.

If B is absent, this causes C's status to match A's: if A is present, this causes C to be present, and if A is absent, this causes C to be absent.

If B is present, this causes C's status to be the opposite of A's: if A is present, this causes C to be absent, and if A is absent, this causes C to be present.

Florence is a mouse from the Muckenhoupt Lab. You test Florence and find that:

Hormone A is definitely present.

Hormone C is definitely not present.

For Florence the mouse, if A were absent, would A be present?

For Florence the mouse, if A were absent, would B be present?

For Florence the mouse, if A were absent, would C be present?

Experiment 4

In almost all of the mice, hormone A is present.

In almost all of the mice, hormone B is present.

The status of hormone C is caused by the status of hormones A and B and nothing else.

If A is present, this always causes C to be present.

If B is present, this always causes C to be present.

The status of hormone D is caused by the status of hormone C and nothing else.

If C is present, this always causes D to be present.

Harold is a mouse from the Weyland Lab. You test Harold and find that:

Hormone A is definitely not present.

Hormone B is definitely not present.

Hormone C is definitely not present.

Hormone D is definitely not present.

For Harold the mouse, if C were present, would A be present?

For Harold the mouse, if C were present, would B be present?

For Harold the mouse, if C were present, would C be present?

For Harold the mouse, if C were present, would D be present?

Experiment 5

In many of the mice, hormone A is present.

In many of the mice, hormone B is present.

The status of hormone C is caused by the status of hormones A and B and nothing else.

If hormone A is present, this always causes hormone C to be present.

If hormone B is present, this almost always causes hormone C to be present.

The presence of hormone D is caused by the presence of hormone C.

If hormone C is present, this always causes hormone D to be present.

Ilana is a mouse from the Snarp Lab. You test Ilana and find that:

Hormone A is definitely present.

Hormone B is definitely present.
Hormone C is definitely present.

Hormone D is definitely present.

For Ilana the mouse, if C were absent, would A be present?

For Ilana the mouse, if C were absent, would B be present?

For Ilana the mouse, if C were absent, would C be present?

For Ilana the mouse, if C were absent, would D be present?

Experiment 6

In some of the mice, hormone A is present.

The status of hormone B is caused by the status of hormone A and nothing else.

In almost all mice, if A is present, this causes B to be present, and if A is absent, this causes B to be absent. In the remaining mice, if A is present, this causes B to be absent, and if A is absent, this causes B to be present.

The status of hormone C is caused by the status of hormone B and nothing else.

In almost all mice, if B is present, this causes C to be present, and if B is absent, this causes C to be absent. In the remaining mice, if B is present, this causes C to be absent, and if B is absent, this causes C to be present.

Albert is a mouse from the Wrean Lab. You test Albert and find that:

Hormone A is definitely not present.

Hormone B is definitely not present.

Hormone C is definitely present.

For Albert the mouse, if you had observed B to be present, would A be present?

For Albert the mouse, if you had observed B to be present, would B be present?

For Albert the mouse, if you had observed B to be present, would C be present?